Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning

Shresth Verma
Google Research India
vermashresth@google.com

Aditya Mate
Harvard University
aditya_mate@g.harvard.edu

Kai Wang
Harvard University
kaiwang@g.harvard.edu

Neha Madhiwalla
ARMMAN
neha@armman.org

Aparna Hegde
ARMMAN
aparnahegde@armman.org

Aparna Taneja
Google Research India
aparnataneja@google.com

Milind Tambe
Google Research India
milindtambe@google.com

ABSTRACT

Mobile Health Awareness programs in underserved communities often suffer from diminishing engagement over time and health workers have to make live service calls to encourage beneficiaries' participation. Owing to health workers' limited availability, we consider the optimization problem of scheduling live service calls in a Maternal and Child Health Awareness Program and model it using Restless Multi-Armed Bandits (RMAB). Since the parameters of the RMAB formulation are unknown, a model is learnt to first predict the parameters of the RMAB problem, which is subsequently solved using the Whittle Index algorithm. However, this Predict-then-Optimize framework maximises for the predictive accuracy rather than the quality of the final solution. Decision Focused Learning (DFL) solves this mismatch by integrating the optimization problem in the learning pipeline. Previous works have only shown the applicability of DFL in simulation setting. In collaboration with an NGO, we conduct a large-scale field study consisting of 9000 beneficiaries for 6 weeks and track key engagement metrics in a mobile health awareness program. To the best of our knowledge this is the first real-world study involving Decision Focused Learning. We demonstrate that beneficiaries in the DFL group experience statistically significant reductions in cumulative engagement drop, while those in the Predict-then-Optimize group do not. This establishes the practicality of use of decision focused learning for real world problems. We also demonstrate that DFL learns a better decision boundary between the RMAB actions, and strategically predicts parameters for arms which contribute most to the final decision outcome.

KEYWORDS

Restless Multi-Armed Bandits; Decision Focused Learning; Population Health

ACM Reference Format:
Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. 2023. Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

1 INTRODUCTION

States Covered in India	19
Partner NGOs	40
Partner Hospitals	97
Health Workers Trained	235K
Beneficiaries	27.2M

Scale of ARMMAN

Figure 1: Beneficiary receiving preventive health information

Non-profits often leverage the extensive cell phone coverage to feasibly reach underserved communities for information dissemination programs. In particular, NGOs working in the mobile health space can deliver timely and targeted health information via text or voice messages [12, 20]. Unfortunately, such programs suffer from a dwindling engagement over time, with large number of beneficiaries dropping out from the program. NGOs can make use of health workers to personally reach out to beneficiaries through service calls, encourage their participation and address complaints. However, health workers' availability and time are scarce resources; only a limited number of beneficiaries can be given a service call every week. It is thus crucial to optimize which beneficiaries receive these personal service calls. We pose this as optimization problem of constrained sequential resource allocation solved using Restless Multi-Armed Bandits (RMAB). Each beneficiary is modelled as an arm following a Markov Decision Process and the action of whether to place a service call or not results in state change. The Whittle index heuristic [30] is the dominant approach for solving RMABs. However, for computing Whittle Indices, transition dynamics of each arm must be known. While many previous works make the assumption that transition dynamics parameters are already known, in the real world, these parameters must be inferred. When arm features are correlated with transition dynamics, historical data

on arm pulls is leveraged to learn a mapping from arm features to transition dynamics [16, 23]. The learnt mapping function is then used to predict the unknown parameters for new arms and solve the subsequent optimization problem.

This approach thus falls under the Predict-then-Optimize [4–6] framework, where an optimization problem is to be solved but the parameters defining the optimization problem are unknown. This is a *two-stage approach*: The first stage is to learn a predictive model which maps from some environment features to the parameters. Subsequently, in the second stage, the optimization problem formulated using the predicted parameters is solved. However, there is a key shortcoming in this two-stage framework. While the mapping function maximizes for the predictive accuracy of parameters, we are interested in the solution quality of the optimization problem parameterized by the predicted parameters. Decision-Focused Learning (DFL) [3, 14, 27, 31] is proposed to address this mismatch between the training objective and the evaluation objective by embedding the optimization problem within the training pipeline. However, until now, Decision Focused Learning has only been studied through simulated experiments.

In this paper, we present the first work showcasing the real-world impact of DFL for RMABs through a large scale field study. For conducting the field study, we collaborate with ARMMAN, an NGO in India working in mobile health space for maternal and child health awareness (Figure 1). In prior works, a RMAB model using the previously mentioned two-stage learning approach has been used for optimizing live service call scheduling in the field [16]. We compare this two-stage approach with a DFL approach in optimizing service calls. Engagement is a key metric that captures beneficiaries' participation in the mobile program. Our results show that allocating health worker resources using a DFL policy reduces drop in engagement by 31% as compared to the no-service call baseline. On the other hand, the benefit from TS policy is not statistically significant. We also show that live service calls made by health care workers using DFL policy have higher effectiveness than TS policy resulting in better short-term as well as long-term outcomes in listenership behaviour.

Furthermore, we perform detailed post-hoc analysis of the real-world study and back the observations using simulated experiments to explain how DFL is making decisions and why those decisions result in a better performance. Our novel contributions are as follows:

- We show results from the first large-scale field study of Decision Focused Learning being applied to maternal and child health domain.
- We show that by optimizing for decision quality rather than predictive accuracy, DFL results in statistically significant improvement in final decision quality measured through engagement metric in the mobile health program.
- We provide an interpretation of how DFL strategically learns to distinguish between arms that benefit most from interventions, resulting in improved parameter predictions compared to the TS model.

Our positive results thus pave the way for future works applying Decision Focused Learning in real world agent-modelling tasks as well as optimization problems with unknown underlying problem parameters. We shall release the code for experiments upon acceptance.

2 RELATED WORK

The optimization problem of constrained sequential resource allocation can be solved using Restless Multi-Armed Bandits (RMAB). RMABs have been used in real world applications such as anti-poaching patrol planning [22], healthcare interventions [15, 16], and machine repair and maintenance [9]. The complexity of optimally solving RMAB problems is known to be PSPACE hard [17]. Whittle Index approach [30] is an approximate solution to RMAB problem which is aymptotically optimal under the indexability condition [1, 28, 29]. However, for computing the Whittle Index, transitions dynamics must be known. Under unknown system dynamics, [16, 23] leverage the predict-then-optimize framework for learning a predictive model of transition dynamics from features using historical data.

The predict-then-optimization [7] framework (or two-stage learning) solves for an optimization problem with unknown parameters by learning a predictive model of parameters from environment features and subsequently solving the optimization problem. However, this two-stage process separates out the prediction and optimization problems, thereby causing a mismatch between the predictive loss that is minimized and the evaluation metric that is desired to be maximized [10, 11, 13]. Decision Focused Learning [6, 14, 31], solves this problem by embedding optimization problem as a differentiable layer in a deep learning pipeline. Most previous DFL [3, 6, 14, 19] approaches solve one-shot optimization problems such as stochastic programming and security games in an end-to-end manner. Recently, [8, 26] propose an extension of Decision Focused Learning for sequential decision making problems. At AAMAS, Decision Focused Learning has been applied in directly optimizing game utilities in Network Security Games [25] and Stackelberg Security Games [18]. [27] further extend the Decision Focused learning methodology for Restless Multi Armed Bandit problems for generalized N-state MDP as well as a belief state MDP to optimze for decision quality. However, none of these works, either in the single shot setting or the sequential decision making settings, have ever been tested in the real world in the field; and hence were unable to thoroughly analyze comparative advantages of decision focused learning over baseline approaches with real world data.

3 MOBILE HEALTH ADHERENCE

3.1 Mobile Health Program

ARMMAN is a non-governmental organization in India focused on reducing maternal and neonatal mortality among underprivileged communities. The NGO operates a mobile health service that disseminates preventive health information to expectant or new mothers (beneficiaries) on a weekly basis via automated voice messages. A large fraction ($\sim$ 90%) of mothers in the program are below the World Bank international poverty line [32] and the program has so far served over a million mothers. However, despite the success of the program, beneficiaries' engagement with the voice calls dwindles over time with 22% of beneficiaries dropping out of

the program within just 3 months of enrolment. Live Service calls made by health workers can encourage beneficiaires' participation. However, the health workers' availability is limited and thus, only a fixed number of live service calls can be made every week. This constraint necessitates a smart scheduling strategy of which beneficiaries to reach out every week to best utilize health workers' efforts.

3.2 Restless Multi-Armed Bandits

We consider the Restless Multi-Armed Bandit model with N independent arms each characterized by a 2-action Markov Decision Process (MDP) Figure 2. Each MDP is defined using the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$ where $\mathcal{S}$ refers to the state space, $\mathcal{A}$ is the action space, which in our case is discrete and binary, $\mathcal{A} \in \{0, 1\}$. R is the reward function such that $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$. $\mathcal{P}$ is the transition function, such that $\mathcal{P}(s, a, s')$, $(s, s') \in \mathcal{S}, a \in \mathcal{A}$ represents the probability of transitioning from state s to s' under action a. The policy function $\pi : \mathcal{S} \mapsto \mathcal{A}$ is defined as the mapping from states to action.

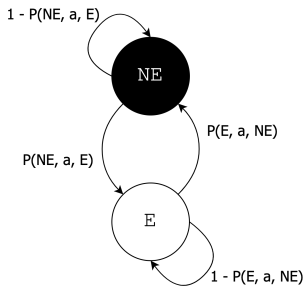


Figure 2: The beneficiary transitions from a current state s to a next state s' under action α, with probability $P(s, \alpha, s')$.

In our problem setup, we consider a 2-state 2-action MDP problem. Based on our discussions with the NGO, states are defined using the engagement metric. If a beneficiary listens to at least 1 call for more than 30 seconds in a week, they are said to be in Engaging state ($s = 1$). Otherwise, the beneficiary is in Non-Engaging state ($s = 0$). The timestep of the MDP is chosen to be a period of 1 week. The actions correspond to whether to deliver (active) or not deliver (passive) live service call to a beneficiary. Additionally, the NGO can only deliver K live service calls in a week. The reward function at any given timestep is defined to be same as the current state $R(s, a) = s$. The planner's goal is then to maximize expected long term reward (engagement). Starting from a state s_0, this is defined using the value function V as :

$$V(s_0) = \mathbb{E}_{s_{t+1} \sim P}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1} | \pi, s_0)\right] \tag{1}$$

where γ is the discount factor for rewards.

The *Whittle Index* for every arm is defined using the 'passive subsidy'. The passive subsidy is the additional reward accrued by an arm when the passive action is chosen. The whittle index is then defined as the passive subsidy such that expected future value is identical for both the passive and active actions. Formally, the whittle index WI_i for an arm i in state s can be defined as:

$$W_i(s) = \inf_m\{V_i^m(s; a = 0) = V_i^m(s; a = 1)\} \tag{2}$$

where V_i^m is subsidized value function under passive subsidy m.

Intuitively, the Whittle index measures the value of pulling an arm conditioned on the observed state. Therefore, at every timestep, the *Whittle Index Policy* ranks all arms by their current state whittle index. The top-K arms with the highest whittle indices are chosen for active action to maximize the total pulling performance.

3.3 Missing Transition Probabilities in RMAB

Most works using RMABs make the assumption that MDP parameters are known beforehand. However, in practice, we may not have access to beneficiaries' transition probabilities to define the RMAB model. In our problem, the mobile health program receives new sets of beneficiaries without information about their transition behavior. This prevents us from applying techniques in RMAB to properly schedule service calls.

Learning challenge. The solution we adopt here is to learn a mapping from the beneficiaries' demographic features and prior interaction with the program to forecast the transition probabilities. Similar to Predict-then-Optimize framework [5] we learn a predictive model and then determine the live service call schedule using the RMAB model.

Dataset. We use the historical beneficiaries' listenership behaviour between January 2022 to May 2022 as the training dataset. Specifically, we have access to state trajectories of 19944 (N) beneficiaries over a period of 5 weeks (T), along with the action chosen for every beneficiary at every timestep. Note that passive actions make up majority of the historical data with only 3% of transitions happening under an active action. In addition to the trajectories, we have socio-demographic features for every beneficiary obtained at registration time. These features cover information such as age, gestational age, income, education, parity, gravidity, language of automated call, and registration channel.

4 COMPARISON OF LEARNING METHODS

In this section, we summarize the Two-Stage and the Decision-Focused learning approaches for obtaining the transition probability parameters of beneficiaries. Crucially, the TS approach maximizes for the predictive accuracy while the DFL approaches maximizes the decision objective.

4.1 Two-stage Learning

In [16], TS model is shown to cut $\sim$ 28% engagement drops as compared to a Round-Robin baseline. In our work, we consider outperforming the TS baseline to show applicability of DFL model. Thus we follow similar setup of the TS model as described in [16]. A mapping function f is learnt that predicts the Transition Probabilities given the socio-demographic features x_i for the i_{th} arm. Predicted Transition Probabilities for arm P_i can then be obtained as $P_i = f(x^i), i \in [N]$. Since our problem domain consists of two states and two actions, we have to predict four transition probabilities. We model the mapping function as a neural network f_w parameterized by the weights w. f_w is designed using two fully

connected layers followed by four outputs and finally logistic function is applied to obtain probabilities. f_w is learnt by minimizing the negative log-liklihood of observed trajectories $\mathcal{T}$ under the predicted transition probabilities $f_w(x)$. The loss function $\mathcal{L}$ is thus given by

$$\mathcal{L}(f_w(x), \mathcal{T}) = \mathbb{E}_{i \in [N]} -\log(\mathcal{T}^i | f_w(x^i)) \tag{3}$$

The weights w of the neural network f_w are optimized by backpropogating the gradient $\frac{d\mathcal{L}(f_w(x),\mathcal{T})}{dw}$.

4.2 Decision-focused Learning

We replicate the Decision Focused learning pipeline from [27] where instead of optimizing for predictive accuracy, the final decision outcome is optimized. Off-Policy Policy Evaluation (OPE) is used to quantify the decision outcome. It measures the reward obtained from a learnt policy given the past trajectories from a different policy. The DFL architecture uses the same predictive model f_w as TS, described in the previous section. However, once Transition Probabilities are predicted as $P = f_w(x)$, we compute Whittle Indices using a differentiable function W. The whittle indices $WI = W(P)$ parameterize a differentiable policy which we denote as π^{WI}. Finally, the differentiable evaluation objective is formulated using OPE of learnt policy under the observed trajectories $\mathcal{T}$ which is represented as $OPE(\pi^{WI}, \mathcal{T})$. The weights of the predictive model are learnt by maximizing the final objective and backpropogating through the complete pipeline. The gradient is thus given by $\frac{d\ \mathrm{OPE}(\pi^{WI},\mathcal{T})}{dw}$.

In Decision Focused Learning, we calculate this gradient by using the chain rule:

$$\frac{d\ \mathrm{OPE}(\pi^{WI}, \mathcal{T})}{dw} = \frac{d\ \mathrm{OPE}(\pi^{WI}, \mathcal{T})}{d\pi^{WI}} \frac{d\pi^{WI}}{dWI} \frac{dWI}{dP} \frac{dP}{dw} \tag{4}$$

We refer the reader to the appendix for more details on DFL pipeline.

5 FIELD STUDY

We collaborated with the NGO on the maternal and child health problem and conducted a service quality improvement field study to compare the performance of different learning approaches. All experiments reported in this paper are approved by an ethics review board at the NGO.

Hypothesis and research question: The main goal in this paper is to understand the performance of decision-focused learning in real-world problems. Decision-focused learning has shown better performance in many applications but only in simulation. There is no deployment or real-world evidences of whether decision-focused learning actually outperforms other learning methods in practice.

Control methods. In earlier work [16], the two-stage approach was shown to outperform a benchmark of Round Robin Policy. The work also provides statistical significance results, illustrating the superiority of two-stage RMAB policy over non-AI baseline. Therefore outperforming the two-stage approach is important to show the utility of decision-focused learning. In our field study, we compare the following live service call scheduling strategies: (1) Current Standard of care (CSOC), where no live service calls are delivered to the beneficiaries, (ii) Two-stage (TS) approach where beneficiaries are chosen for live service calls according to the Whittle Index Policy learnt using Two-Stage learning, and (iii) Decision-Focused Learning (DFL) approach where beneficiaries are chosen for live service calls according to the Whittle Index Policy learnt using Decision Focused learning. We use the performance of the CSOC group to anchor the performance of other AI-based methods. The performance of the CSOC group also measures the baseline engagement rate that the mobile health program receives without any intervention. Therefore, we focus on the improvement of AI-based methods against the CSOC method.

Eligibility criterion and randomization. We consider the group of beneficiaries registered between the months of April 2022 to June 2022. Further, we filter out beneficiaries who have not listened to even a single automated voice call in the time period of 30 days before the study begins. This filtering is done to remove beneficiaries from the cohort who have long term connectivity issues such as phone number out of service or misentry of phone number at enrolment. Lastly, we randomly sample 9000 beneficiaries out of these eligible candidates to form our study cohort. We split these set of beneficiaries into three groups of 3000 beneficiares each - (i) CSOC group, (ii) TS group, and (iii) DFL group. We make sure that the distribution of socio-demographic features and start-state are the same across the three groups.

Experiment design. Beneficiaries become eligible for live service calls 2 months post their enrolment into the program. Within the TS and DFL groups, we choose $K = 300$ beneficiaries for live service call every week based on NGO's constraints. These live service calls are sent out weekly for a period of 6 weeks. We continue to monitor the cohort for 4 more weeks even after the study ends to measure the sustained effect of live service calls. It should be noted that, automated voice messages are sent to all groups throughout this period and only the delivery of live service calls by health workers changes across the three groups.

6 EXPERIMENT RESULTS

In this section, we showcase the results from the field study. We also define multiple evaluation metrics and demonstrate how the different policies fare against each other.

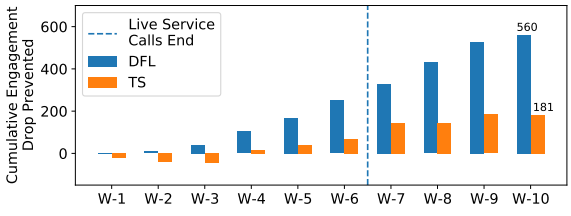


Figure 3: Weekly Cumulative Engagement Drop Prevented for the DFL and TS groups. Live service calls are only delivered for the first 6 weeks, after which, all three groups are only passively observed. The DFL group prevents more Cumulative Engagement Drops as compared to the TS group

6.1 Weekly and Cumulative Engagement

We first present the results from our study using the Engagement Metrics proposed by Mate et al. [16]. Engagement at time t for the i^{th} beneficiary, represented by $E^i(t)$, is defined as 1 if the beneficiary listens to at least one automated call in a week for more than 30 seconds and 0 otherwise. Since the engagement of beneficiaries dwindles over time, we can measure the drop in engagement relative to the engagement at start. The engagement drop and the cumulative engagement drop are defined as

$$E^i_{drop}(t) := E^i(0) - E^i(t);\ E^i_{cumu\_drop}(t) := \sum_{\zeta=0}^{\zeta=t} E^i_{drop}(\zeta)\ . \quad (5)$$

The cumulative engagement drop prevented over the CSOC group is simply the difference in cumulative engagement drop of the policy and the CSOC group. Figure 3 shows the cumulative engagement drops prevented over CSOC group for DFL and TS policies. We see that DFL prevented more drops across all weeks and by the end of study, DFL group has **560** more engagement drops prevented over the CSOC group as compared to TS group which only prevents **181** engagement drops. Given a total of 1765 cumulative engagement drops in the CSOC group, DFL group has 31% fewer cumulative engagement drops as compared to CSOC group while TS only results in 10% reduction in cumulative engagement drops.

6.2 Statistical Significance

We also establish statistical significance [1] of DFL's benefit using regression analysis [2]. We fit a linear regression model to predict the output variable $E^i_{cumu\_drop}$ by giving beneficiary features x_i as an input vector of length J along with and an indicator variable T_i denoting whether a beneficiary belongs to DFL ($T_i = 1$) or CSOC ($T_i = 0$) group. The regression model can thus be represented as

$$Y_i = k + \beta T_i + \sum_{j=1}^{J} \gamma_j x_{i,j} + \epsilon_i \quad (6)$$

where β is the regression coefficient of the indicator variable T_i measuring the effect of treatment, γ_j is the regression coefficient of the j-th input feature, k is the constant term of regression and ϵ_i is the error. Y_i is the target variable that is fitted using the regression model and is same as $E^i_{cumu\_drop}$. The regression coefficient for T is found to be 0.19 with p-value of 0.024. On the other hand, similar comparison between TS ($T_i = 1$) vs CSOC ($T_i = 0$) yields a regression coefficient of 0.06 for T with p-value of 0.48. Thus, belonging to the DFL group resulted in significantly positive impact on cumulative engagement drops while for TS, no such statistical significance could be established.

6.3 Performance on Listenership Metrics

While the whittle index policy maximimizes the reward, which is defined using the engagement metric, we also measure if the policy improved other metrics characterizing listenership. Thus, we define metrics quantifying listenership behaviour of a beneficiary within a time window of 14 days before and after receiving a service call.

[1] See Appendix **??** for erratum

Table 1: Statistical significance for service call impact tested using a linear regression model

	DFL vs CSOC	**TS vs CSOC**
% reduction in cumulative engagement drops	31%	10%
p-value	0.024	0.48
Coefficient β	0.19	0.06

Table 2: Performance of the DFL and TS policies across multiple listenership metrics. DFL policy shows a higher change in listenership behaviour from a service call as compared to the TS policy.

Policy	**Change in Mean Duration**	**Change in No. of Engagements**	**Change in E/S**
DFL	17.054	0.094	0.20
TS	6.764	0.009	0.07

Definitions.

(1) Mean Duration: The mean duration of calls listened to within the time window.
(2) No. of Engagements: The numbers of calls engaged with (30+ seconds listened) within the time window.
(3) Engagements to Scheduled (E/S) Ratio: The ratio of numbers of calls engaged with to numbers of calls scheduled within the time window.

Results. We calculate the change in these metrics between the time window before and after a live service call. Table 2 reports the mean change in these metrics across the three experimental groups. We observe that across all the metrics, **DFL group has a significantly higher change in listenership behaviour through live service calls as compared to the TS group**. For instance, we can interpret the value of 17.054 in Mean Duration metric for DFL as active actions in DFL group resulting in beneficiaries listening to on average 17 seconds more of an automated call. This is in contrast to TS group, where live service calls only resulted in beneficiaries listening to 6 seconds more of an automated call. Note that the average duration of an automated message is 60 seconds. Thus a 17 seconds improvement in listenership corresponds to an average 28% increase in message content listened to among those treated with live service calls. Using t-test for comparison of means, we find that for each of the proposed metrics, mean change is statistically higher for DFL group as compared to TS group with p-value< 0.05.

7 UNDERSTANDING DFL

7.1 Learnings from Real World Experiment

The Decision Focused Learning method consists of an end-to-end pipeline starting from features to predicted Transition Probabilities to computed whittle index and finally the decision of whether a beneficiary is in top-K list chosen for live service call. In this section, we interpret the DFL's strategy in contrast with the Two-Stage policy by performing post-hoc analysis across all these steps.

Table 3: Multiple Error and Rank metrics evaluated for DFL and TS policies. While TS group shows a lower overall error in predicting transition probabilities, DFL group has lower predictive error in Top-K arms and a higher rank correlation with the optimal ranking.

	Rank Metrics		Transition Probability Error Metrics			
Policy	Precision @ K	Spearman's Correlation	MAE All	MAE Top-K	Mean NLL All	Mean NLL Top-K
DFL	0.41	**0.30**	0.31	**0.35**	0.79	**0.62**
TS	0.22	0.179	**0.25**	0.37	**0.42**	0.69

As a first step for this analysis, we compute the ground truth transition probabilities using the observed trajectories of beneficiaries during the time period of field study. Once Ground Truth Transition Probabilities are estimated, we subsequently compute the Ground Truth Whittle Index and Ground Truth top-K ranks.

Top-K Rank Lists. We consider the ordered list of beneficiaries according to predicted whittle index in the Two-Stage and DFL experiment groups. Additionally, True Top-K rank list is also computed using the ground truth whittle index. To measure the agreement between the two lists, we use the following metrics:

(1) Precision @ K: This metric counts the proportion of relevant beneficiaries in the top-K positions of the policy rank list and is widely used in classification [21, 33] and ranking problems [24]. The precision @ K in our problem is given by:

$$\text{Precision @ K} = \frac{|\text{Policy Top-K list} \cap \text{True Top-K list}|}{K}$$

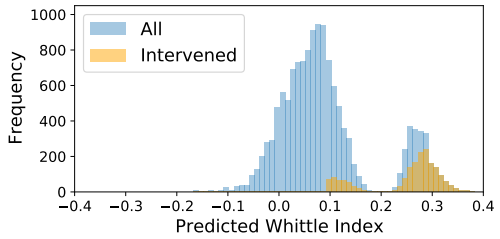
(2) Spearman's Rank Correlation: This metric calculates the rank correlation between the Predicted Whittle Index and Ground Truth Whittle Index of Policy's Top-K ranked beneficiaries.

In Table 3, we show the different rank metrics for the two comparison groups. In all the weeks, we find that **the DFL group has a higher agreement with the True Top-K ranks as compared to the Two-Stage experiment group.**

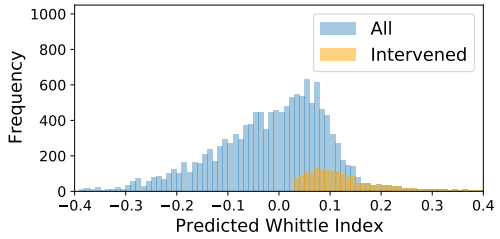
Whittle Indices. For beneficiaries belonging to each of the experimental group, we have the corresponding computed Whittle Index from predicted Transition Probabilities. We call it the Predicted Whittle Index (note that these values are not directly predicted by the Neural Network models). Figure 4 shows the distribution of Predicted Whittle Index for DFL and TS experiment groups in Blue. We also mark the beneficiaries who are chosen for Active action within each experimental group in orange.

A striking observation is that **within the DFL group, the whittle indices have a bimodal distribution as opposed to a unimodal distribution for Two-Stage group**. This suggests that in DFL, the model is trying to learn a decision boundary between the beneficiaries to deliver active and passive action. This contrasts with the Two-Stage model where objective is solely to learn accurate transition probabilities.

Predicted Transition Probabilities. Given the ground truth and predicted transition probabilities for both DFL and TS policies, we compute for the whole population (i) the Mean Negative Log Likelihood (NLL) of observed trajectories under predicted transition probabilities and (ii) the prediction error using Mean Absolute Error (MAE). In Table 3, we show that DFL has both higher MAE and



(a) Decision Focused Learning



(b) Two Stage

Figure 4: Predicted Whittle Index Distribution and Beneficiaries Intervened for TS and DFL groups across all weeks. The DFL group has a bimodal distribution of predicted whittle index as compared to unimodal distribution in the TS group. Note that the right peak in DFL is not fully covered due to beneficiaries changing states over the course of study.

higher Mean NLL as compared to TS. Thus DFL model is poorer in predicting the transition probabilities. However, if we compute these metrics for just the true top-K beneficiaries (MAE Top-K and Mean NLL Top-K), we find that DFL has lower MAE as well as Mean NLL than TS. This suggests that the **DFL focuses on correctly predicting the transition probabilities for beneficiaries who will actually be picked, in contrast to the TS, which optimizes for predictive performance for the whole population**. It must be noted, that the predictive performance of TS is impacted due to limited historical data around active actions (limited service calls made by the NGO).

7.2 Short-term and Long-term Impact of Live Service Calls

In Figure 5, we plot the mean reward accrued by beneficiaries in the immediate next step after an active action for both the Two-Stage and the DFL group. This quantifies the short term impact of a live

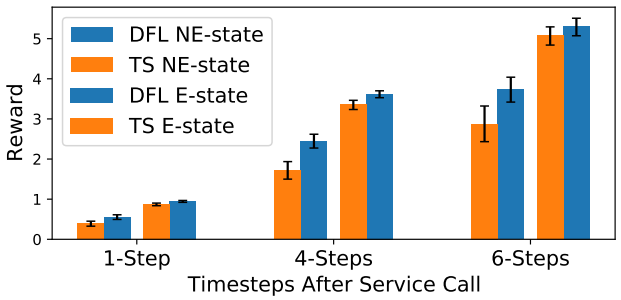


Figure 5: Mean reward accrued by beneficiaries in short term (1-step lookahead reward) and long term (4-steps and 6-steps lookahead rewards) after given an active action, DFL group has higher reward in both the short-term and the long term as compared to the TS group.

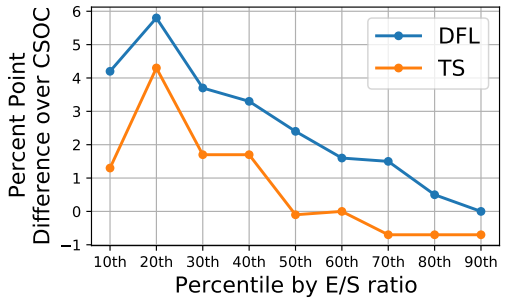


Figure 6: Lift in E/S ratio over CSOC for different percentiles. The highest lift in E/S ratio is in the lowest percentile suggesting that beneficiaries with poor listenership of automated voice messages benefited the most from live service calls.

service call. In both the NE and E state, we observe that DFL leads to higher reward.

While short-term impact is only applicable for one timestep ahead, the Whittle Index policy optimizes for long-term rewards. In Figure 5, we also plot the reward obtained in 4 weeks and 6 weeks following the live service call. We show this for both TS and DFL group. Again, we see that DFL's live service calls are more effective than TS policy even in the long term.

7.3 Who Benefits from DFL

In order to determine which beneficiaries benefited the most from the DFL policy, we first obtain the ratio of calls engaged with over total scheduled calls (E/S) for every beneficiary over the whole duration of study. Subsequently, we rank the beneficiaries based on the E/S ratio and compute average E/S ratio for different percentiles. We calculate these numbers for all three policies. In Figure 6, we plot the lift in E/S ratio over CSOC for different percentiles. While DFL shows a positive lift in listenership over CSOC across all percentiles, the maximum lift is achieved in the lowest percentiles. This shows that those with low listenership are the ones benefiting most from the DFL policy.

7.4 Learnings from Simulated Experiments

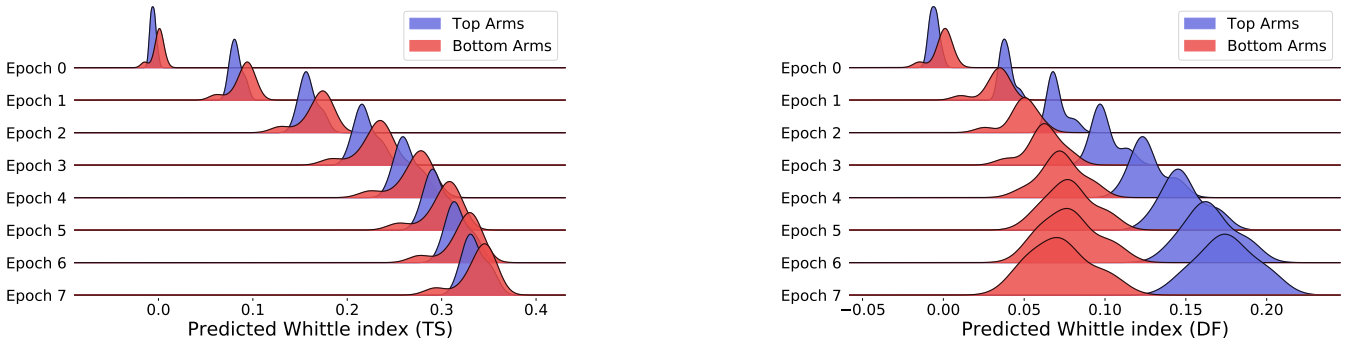
In this section, we conduct simulated experiments to improve our understanding of the DFL model and verify the observations made from the real world experiment. Specifically, we consider an RMAB system with 100 arms simulating beneficiaries enrolled in the NGO's program. The MDP parameters of each arm are randomly initialized. Additionally, we obtain a feature vector corresponding to every arm such that the features are correlated with the MDP parameters. Lastly, we simulated multiple trajectories for the whole system and store that as offline dataset for our experiments. All experimental results are reported by averaging over five seed values.

The Effect of Training Data Size. While Decision Focused Learning optimizes for the decision objective, a TS model that perfectly predicts the optimization problem parameters should also achieve the optimal decision objective. However, in the real world, predictive models do make errors. These errors can be dependent on the quantity of training data that is available to the learning model.

We thus formulate the hypothesis that the gain from a DFL model should be higher in limited data scenario. As the size of training data grows, DFL and TS should converge to similar decision objective. To test this hypothesis, we run a simulated experiment with varying number of trajectories per arm. Figure 8a shows the lift in Off-Policy Policy Evaluation from DFL over TS with increasing training data size. We observe that the highest lift is with smallest training data size and as we increase the availability of training data, the lift diminishes.

Shift in Whittle Index Distribution over Training Epochs. As DFL learns to optimize the decision objective directly, we hypothesise that it should learn a model which effectively separates the top ranked and bottom ranked whittle index arms. On the other hand, since TS optimizes for predictive accuracy, it has no incentive to learn an optimal ranking of the arms by whittle index. To verify this hypothesis, we plot the predicted whittle index distributions of true top-K and bottom-K arms. In Figure 7, we visualize how these distributions change over the training epochs, giving a glimpse into the learning process of the two models. We observe that both the TS and DFL model start with no prior information of the true top-K and bottom-K arms. However, over the training epochs, DFL learns whittle indices such that it separate the two groups. The Two-Stage model fails to learn such segregation in predicted whittle index distribution.

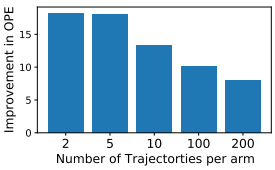
The Effect of Budget-K. The budget constraint in the RMAB problem defines the number of arms chosen for active action every week. In a two-stage model, the learning step outputs the transition probabilities irrespective of the budget value K. However, in Decision Focused Learning, the mapping model which outputs the transition probabilities maximizes for the decision objective that relies on the value of K. To simulate the effect of mismatch in K, we train DFL model with changing K at train time (K_train), while keeping the K fixed at the time of evaluation (K_eval). Specifically, we note the OPE at evaluation time with $K = 20$ for different training scenarios with $K \in [2, 4, 10, 16, 20, 30, 40, 60, 80]$ as shown in Figure 8b. We observe that the performance at evaluation time only drops slightly (by upto 6%) in both the cases of train time budget being greater or lesser than the evaluation time budget. The sensitivity of the



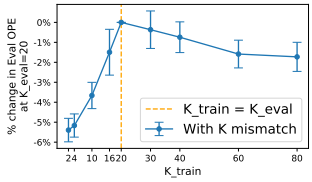
(a) Two-Stage Learning

(b) Decision Focused Learning

Figure 7: Predicted whittle index distribution for optimal top and bottom arms, across the training epochs. DFL policy learns whittle indices such that the true top ranked and bottom ranked arms are segregated. TS policy fails to learn whittle indices following this strategy.



(a) Improvement in Evaluation OPE of DFL over TS with changing number of trajectories per arm.



(b) % Drop in Evaluation OPE with budget as 20 (K_eval) relative to maximum Eval OPE, across changing budget at train time (K_train).

DFL's performance to the value of K_train supports the hypothesis that DFL learns a decision boundary optimized for the exact number of beneficiaries chosen for active action. Further, keeping $K\_eval = K\_train$ can help maximize the performance of DFL.

8 CONCLUSION

Several applications at AAMAS first require learning a predictive model of agents' parameters and then optimizing based on result of such learning. This paper presents key results on importance of Decision Focused Learning to such applications. We conduct the first large-scale field study of Decision Focused learning through an RMAB problem in maternal and child health domain. We conclude that learning the MDP parameters of the RMAB problem through Decision Focused Learning results in higher participation of beneficiaries in the program (Figure 3). DFL's strategic selection of actions also results in more effective live service calls as demonstrated in Table 2. From the analysis showcased in previous sections, we attribute the success of DFL to the following: (i) The predicted whittle index distribution from DFL policy is bimodal in contrast to a unimodal distribution in TS (see Figure 4) indicating that DFL model learns a decision boundary to highly rank beneficiaries that would benefit significantly more from receiving the service call than the rest of the population. (ii) DFL is more aligned with the optimal policy as shown by a higher rank correlation with the True Top-K Beneficiaries as compared to TS (Table 3). (iii) While TS results in a lower predictive error for the population as a whole, DFL by optimizing for decision quality results in improved transition probability prediction for the top-K beneficiaries (Table 3).

9 RESPONSIBLE DEPLOYMENT AND DATA USAGE

We recognize the responsibility associated with deploying real-world AI systems that impacts underserved communities. In our approach, we have iteratively designed, developed and deployed the system in constant coordination with an interdisciplinary team of ARMMAN's field staff, social work researchers, public health researchers and ethical experts. Particularly, all experiments, field tests and the deployment were performed after obtaining approval from ethics review board at both ARMMAN and Google.

Consent and Data Usage. The consent for participating in the mHealth program is received from beneficiaries. Additionally, all the data collected through the program is owned by the NGO and only the NGO is allowed to share data. This dataset will never be used by Google for any commercial purposes. 's data pipeline only uses anonymized data and no personally identifiable information (PII) is made available to the AI models. The data exchange and use was thus regulated through clearly defined exchange protocols including anonymization, read-access only to researchers, restricted use of the data for research purposes only, and approval by ARMMAN's ethics review committee.

Universal Accessibility of Health Information. The mHealth program focuses on improving quality of service calls and does not alter, for any beneficiary, the accessibility of health information. All participants will receive the same weekly health information by automated message regardless of whether they are scheduled to receive service calls or not. The service call program does not withhold any information from the participants nor conduct any experimentation on the health information. The health information is always available to all participants, and participants can always request service calls via a free missed call service.

REFERENCES

[1] Nima Akbarzadeh and Aditya Mahajan. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 7294–7300.

[2] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics*. Princeton university press.

[3] Priya L Donti, Brandon Amos, and J Zico Kolter. 2017. Task-based end-to-end model learning in stochastic optimization. *arXiv preprint arXiv:1703.04529* (2017).

[4] Othman El Balghiti, Adam N Elmachtoub, Paul Grigas, and Ambuj Tewari. 2019. Generalization bounds in the predict-then-optimize framework. *Advances in neural information processing systems* 32 (2019).

[5] AN Elmachtoub and P Grigas. 2017. Smart "Predict, then Optimize". arXiv e-prints. *arXiv preprint arXiv:1710.08005* (2017).

[6] Adam Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. 2020. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*. PMLR, 2858–2867.

[7] Adam N Elmachtoub and Paul Grigas. 2021. Smart "predict, then optimize". *Management Science* (2021).

[8] Joseph Futoma, Michael C Hughes, and Finale Doshi-Velez. 2020. Popcorn: Partially observed prediction constrained reinforcement learning. *arXiv preprint arXiv:2001.04032* (2020).

[9] Kevin D Glazebrook, Diego Ruiz-Hernandez, and Christopher Kirkbride. 2006. Some indexable families of restless bandit problems. *Advances in Applied Probability* 38, 3 (2006), 643–672.

[10] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. 2019. Addressing the loss-metric mismatch with adaptive loss alignment. In *International Conference on Machine Learning*. PMLR, 2891–2900.

[11] Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (2019), 1–54.

[12] Jasvir Kaur, Manmeet Kaur, Venkatesan Chakrapani, Jacqui Webster, Joseph Santos, and Raj Kumar. 2020. Effectiveness of information technology–enabled 'SMART Eating' health promotion intervention: A cluster randomized controlled trial. *PLOS ONE* 15 (01 2020), e0225892. https://doi.org/10.1371/journal.pone.0225892

[13] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. 2020. Objective mismatch in model-based reinforcement learning. *arXiv preprint arXiv:2002.04523* (2020).

[14] Jayanta Mandi, Peter J Stuckey, Tias Guns, et al. 2020. Smart predict-and-optimize for hard combinatorial optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1603–1610.

[15] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Intervention. *Advances in Neural Information Processing Systems* 34 (2020).

[16] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (2022), 12017–12025.

[17] Christos H Papadimitriou and John N Tsitsiklis. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

[18] Andrew Perrault, Bryan Wilder, Eric Ewing, Aditya Mate, Bistra Dilkina, and Milind Tambe. 2019. Decision-focused learning of adversary behavior in security games. *arXiv preprint arXiv:1903.00958* (2019).

[19] Andrew Perrault, Bryan Wilder, Eric Ewing, Aditya Mate, Bistra Dilkina, and Milind Tambe. 2020. End-to-end game-focused learning of adversary behavior in security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1378–1386.

[20] Angela Pfammatter, Bonnie Spring, Nalini Saligram, Raj Davé, Arun Gowda, Linelle Blais, Monika Arora, Harish Ranjani, Om Ganda, Donald Hedeker, Sethu Reddy, and Sandhya Ramalingam. 2016. mHealth Intervention to Improve Diabetes Risk Behaviors in India: A Prospective, Parallel Group Cohort Study. *Journal of Medical Internet Research* 18 (08 2016), e207. https://doi.org/10.2196/jmir.5712

[21] Yashoteja Prabhu and Manik Varma. 2014. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 263–272.

[22] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless Poachers: Handling Exploration-Exploitation Tradeoffs in Security Domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, Catholijn M. Jonker, Stacy Marsella, John Thangarajah, and Karl Tuyls (Eds.). ACM, 123–131.

[23] Yao Sun, Gang Feng, Shuang Qin, and Sanshan Sun. 2018. Cell association with user behavior awareness in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology* 67, 5 (2018), 4589–4601.

[24] Ming Tan, Tian Xia, Lily Guo, and Shaojun Wang. 2013. Direct optimization of ranking measures for learning to rank models. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 856–864.

[25] Kai Wang, Andrew Perrault, Aditya Mate, and Milind Tambe. 2020. Scalable Game-Focused Learning of Adversary Models: Data-to-Decisions in Network Security Games.. In *AAMAS*. 1449–1457.

[26] Kai Wang, Sanket Shah, Haipeng Chen, Andrew Perrault, Finale Doshi-Velez, and Milind Tambe. 2021. Learning MDPs from Features: Predict-Then-Optimize for Sequential Decision Making by Reinforcement Learning. *Advances in Neural Information Processing Systems* 34 (2021).

[27] Kai Wang, Shresth Verma, Aditya Mate, Sanket Shah, Aparna Taneja, Neha Madhiwalla, Aparna Hegde, and Milind Tambe. 2022. Decision-Focused Learning in Restless Multi-Armed Bandits with Application to Maternal and Child Care Domain. *arXiv preprint arXiv:2202.00916* (2022).

[28] Kehao Wang, Jihong Yu, Lin Chen, Pan Zhou, Xiaohu Ge, and Moe Z Win. 2019. Opportunistic scheduling revisited using restless bandits: Indexability and index policy. *IEEE Transactions on Wireless Communications* 18, 10 (2019), 4997–5010.

[29] Richard R Weber and Gideon Weiss. 1990. On an index policy for restless bandits. *Journal of applied probability* (1990), 637–648.

[30] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 25, A (1988), 287–298.

[31] Bryan Wilder, Bistra Dilkina, and Milind Tambe. 2019. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1658–1665.

[32] World Bank. 2020. *Poverty and shared prosperity 2020: Reversals of fortune*. The World Bank.

[33] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 271–278.