Medtronic LABS Final Insights Grant Report

November 14th, 2023

Report Goal: Capture the results of your Accelerator project from a programmatic perspective. We welcome reports that your organization is already drafting for other audiences, so as not to duplicate existing efforts.

1. Project Summary & Background

Our project is focused on predicting a patients' likelihood of enrollment from our screened patient population to drive more equitable access to healthcare. We recognized that some patients screened in the field who were eligible would not travel to a nearby clinic to enroll, or some who are enrolled would not be engaged and get lost to follow up. If we can anticipate who is unlikely to enroll, we foresee creating targeted strategies to customize their patient journey (increased time spent on education, discussion of transportation options, SMS messaging and phone calls etc.) to increase their engagement with the health system. This project will be testing whether patient characteristics at the time of community screening can determine the likelihood of enrollment. Our project has the following goal, objectives, deliverables, and data hypothesis:

- *Goal:* Our goal is to uncover insights that may predict program enrollment in order to design programmatic and product features to enhance health service delivery for underserved patients.
- *Objectives:* Our objective is to strengthen referral conversion by first building a model to predict likelihood of program enrollment and secondly by designing targeted interventions using community health workers (CHWs) and healthcare professionals to outreach to identified patients.
- *Deliverables:* We aim to deliver a predictive model for enrollment, findings from A/B testing targeted interventions, and lessons learned from AI/ML pipeline development.
- *Data Hypothesis:* For community health workers who screen patients for diabetes and hypertension in underserved communities, we want to build a classification model to anticipate which screened patients are unlikely to enroll in health management services at the referral hospital, so that we can help CHWs adjust their time and energy toward patients who may be facing the greatest barriers in accessing healthcare.

Milestones

We broke the project down into 6 main phases; each of which maps to critical milestones for the project. The milestones that we defined as a part of this project are as follows:

1. Project Definition & Scoping
2. Data Cleaning, Preparation & Analysis
3. AI/ML Pipeline Development

4. Modeling, Experimentation & Analysis
5. Product Integration, Interventions & Impact
6. Reporting & Wrap-up

Given that our mid-grant report shared out on learnings from the first half of the grant period, we will begin with the learnings and insights since then. In particular, we begin with milestone 4 (Modeling, Experimentation & Analysis) which was still in progress at the time of our last report. For these milestones, we'll be sharing out the:

1. Process: Context of what was completed during the milestone
2. Insights: What insights were gained during the duration of this milestone?

After walking through each milestone's insights, we will then address the questions of:

3. Human Rights: How might these insights advance the safeguarding of human rights?
4. Future Impact: How will your experience with this data approach serve your team or organization's work in the future?
5. Share-out: How might your use case and learnings be applied to other nonprofits that face similar data challenges?

2.1 Milestone 4: Model Development & Experimentation

Process:

After the data deep dive and cleaning, we began experimenting with various models using the PyCaret library which allows you to train 14 models at once. From this we utilized the top 3 performing models as we experimented with which subset of features to include. After this, we moved into the hyperparameter tuning phase using Optuna, landing on our final set of models to use for deployment.

Insights:

An imbalanced dataset risks overfitting to non-useful attributes. Once we began trying to implement an understanding of time into our model, we quickly realized that a majority of patients who will get enrolled, do so within the first 9 days of being screened. Therefore, when we defined a feature around the number of days since the patient had been screened, the enrolled population had a significantly shorter time spans that has lapsed after screening (see Figure 1 for visualization of these distributions). In contrast, those who have still yet to be enrolled have today's current date as the number of days that have lapsed since screening. When we provided the model with the entirety of the dataset, the model quickly exploited this imbalance in the dataset, making the "Days Lapsed" feature the most important feature. See Table 1.

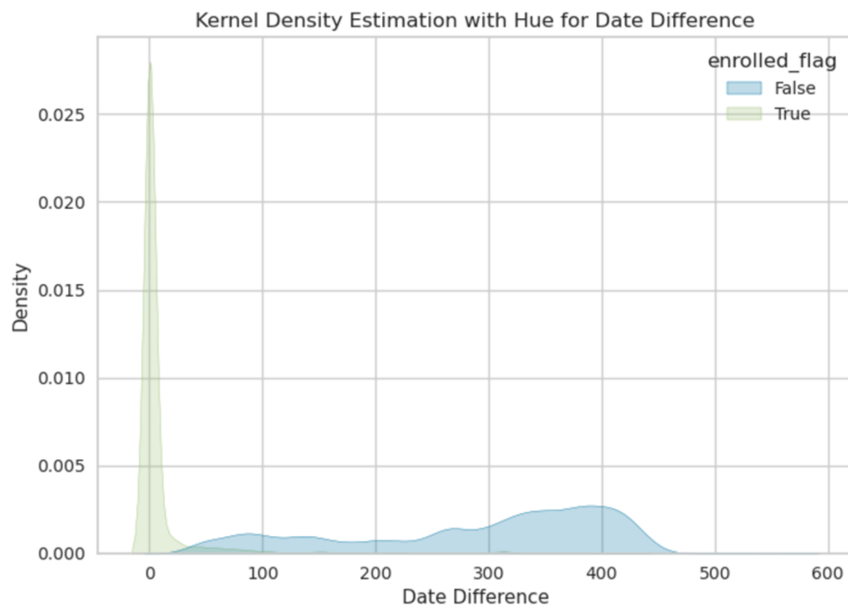


Figure 1. Kernel Density plot demonstrating that those patients who enrolled had significantly fewer days lapsing since they've been screened.

Feature Importance

	feature	feature importance
0	date_difference	0.48
1	bmi	0.12
2	avg_diastolic	0.10
3	avg_systolic	0.06
4	age	0.04
5	country_name_Tanzania	0.04
6	is_before_htn_diagnosis_True	0.04
7	country_name_Sierra Leone	0.04
8	category_Facility	0.02
9	is_regular_smoker_True	0.02
10	is_before_diabetes_diagnosis_True	0.02
11	country_name_Kenya	0.02
12	gender_Non-Binary	0.00
13	gender_Male	0.00
14	cvd_risk_level_Low risk	0.00
15	cvd_risk_level_Medium high risk	0.00
16	cvd_risk_level_Medium risk	0.00
17	cvd_risk_score	0.00
18	type_Inpatient	0.00
19	type_OPD	0.00
20	type_Outpatient	0.00
21	type_camp	0.00
22	type_other	0.00
23	type_pharmacy	0.00

Table 1. The feature importance of the model when trained including our date_difference feature clearly shows that the model exploits this aspect, making most other attributes unique to the patient to be inconsequentially considered.

We also discovered that while we had plenty of patients who had enrolled on the same day that they were screened, we had no "Day 0" patients for those that hadn't enrolled. This is because there is a 1 day lag before completed screening logs to be accessible in our database. Thus, no unenrolled patients are visible from "Day 0". This became a huge bias for the model, as it could easily classify all "Day 0" patients as "Enrolled". Thus, we tentatively decided to remove any patients who enrolled on the same day.

However, by eliminating any patients who enrolled on the same day, this accounted for 70% of the enrolled population (Table 2). Thus, our training data became even further skewed towards those patients who did not enroll.

Data Difference vs Enrolled Patient Counts

Days post screening	Patients enrolled on that day	% of enrolled population
0	17543	0.694470
1	1556	0.061597
2	892	0.035311
3	334	0.013222
4	273	0.010807
7	212	0.008392
5	210	0.008313
8	138	0.005463
6	133	0.005265
9	113	0.004473

Table 2. We discovered that almost 70% of patients who enroll do so on the day they are screened.

To compensate for this large imbalance (almost 31k patients non-enrolled in our dataset and only 7.5k remaining that had enrolled), we down-sampled the unenrolled patients, namely eliminating patients for whom it had been more than 6 months since they had been screened. Now training on a more balanced dataset (5960 enrolled, 8095 not enrolled), the feature importance of time (Table 3) dropped significantly while retaining about 90% specificity for identification of unenrolled patients.

Feature Importance

	feature	feature importance
0	date_difference	0.238667
1	bmi	0.118000
2	avg_systolic	0.114333
3	avg_diastolic	0.112667
4	age	0.106667
5	cvd_risk_score	0.048000
6	country_name_Sierra Leone	0.042000
7	is_before_htn_diagnosis_True	0.030333
8	country_name_Tanzania	0.030000
9	type_camp	0.029000
10	country_name_Kenya	0.029000
11	category_Facility	0.026333
12	type_Outpatient	0.020667
13	is_regular_smoker_True	0.010667
14	is_before_diabetes_diagnosis_True	0.010000
15	type_OPD	0.009667

Table 3. Updated, final feature importance of our model

Our location data had significant gaps and quality issues. Not all screening logs had a GPS location tagged and thus we had to do our best to come up with an alternative understanding of where the patient was located in the community. This proved futile as our "distance" parameter eventually proved to be unimportant in the model's decision-making. Therefore, we learned that if we want to explore further the role that distance and location could play in patient engagement prediction, we need to roll-out a more robust, accurate means of collecting the location data. We also noticed that the site attributes were missing for few sites, prompting us to ask the telecounselors to fill in the gaps, and create a model non-dependent on site features in the interim.

From this early learning, we've now prioritized a Mobile Device Management (MDM) technology to allow location services and network to always be turned on for our devices. We are also working with the MDM provider to collect location offline using the phone's innate GPS hardware. This gap identified early-on has allowed us to prepare for this parameter to be robust as we scale across Kenya and Sierra Leone.

Multiple models can be helpful to compensate for data gaps: As new facilities are added to our programs, we don't immediately have the telecouseling features available for that facility since calls have yet to be placed to those patients. Thus, we needed to create two models, one that utilizes the telecounseling facility features, and one that doesn't require them. Once sufficient answers have been obtained by the telecounseling calls for a new facility, patients from these facilities will have their enrollment prediction score updated by instead being processed through the second model which utilizes these site features.

In addition, as we implement in different countries, there may be different gaps in the data between countries (some questions are mandatory in countries but not others). By creating a new model trained on both Africa and Bangladesh data but only including the features mandatory in Bangladesh, we didn't have to eliminate the Bangladesh patients that had the many of the features missing that were mandatory for the larger African model.

Transferability of our model: Our model in Africa proved to be useful in improving enrollment prediction results in Bangladesh. We had about 2 weeks remaining in the grant and decided to see how we could extend our learnings to a recently kicked off implementation of SPICE in Bangladesh. Because we haven't made many telecounseling calls and the programs are quite young there, we decided to try training a separate model first using only the Bangladesh patients, and then by mixing the patients in with the African population by removing telecounseling features. Training only with the Bangladesh patients achieved us about 81% specificity in identifying patients who would not enroll. We then decided to try mixing Bangladeshi patients in with the African population in order to boost the size of the dataset (13k Bangladeshi patients -> 130k patients combined with African data). This improved the performance to 97% specificity for Bangladeshi patients who wouldn't enroll, and sensitivity increased from 80% -> 94%. Insight here is that there are patterns across our implementations in Bangladesh and Africa such that despite our small dataset in Bangladesh, we can utilize on our large dataset in Africa to contribute to a model tuned for performing well on the Bangladesh data.

Feature importance of our Bangladesh model: By shuffling in our Bangladesh patients, the most important features significantly changed. Shown in Table 4, time since screening still prevailed as the most important feature, but the next 7 features (except for BMI) are programmatic features. This insight demonstrates that a patient's likelihood of enrolling is much more dependent on the health system's history of health service delivery, rather than the patient's individual attributes (although the blood pressures, age and CVD risk scores follow right behind the programmatic features).

	feature	feature importance
0	date_difference	0.204667
1	life_improved_ratio	0.083333
2	enroll_ratio	0.081333
3	actively_managed_ratio	0.063667
4	bg_count_3_mnths	0.060000
5	bmi	0.058333
6	bp_count_3_mnths	0.053000
7	user_count	0.051000
8	avg_systolic	0.049333
9	avg_diastolic	0.046667
10	age	0.041667
11	educational_index	0.027667
12	cvd_risk_score	0.026000
13	health_index	0.024333
14	income_index	0.024000
15	life_expectancy	0.021000

Table 4. The final feature importance of our best-performing Africa-Bangladesh enrollment prediction model.

Explainability of our model: To unpack which attributes were contributing and how (I.e. positive or negative influence on likelihood to enroll), we visualized the SHAP values associated with each feature in our model. Some relationships that are interesting to note and could be of value for other organizations looking to engage patients in the health system. Visualized in **Figure 2** are the following points:

- If patients from a facility generally denote that they utilize the facility as their means of monitoring their blood pressure, that positively correlates with patients' likelihood to follow-through on their screening referral. Conversely, if they mention going elsewhere to check their pressures, this negatively influences patients' likelihood to follow through on enrollment from screening. This indicates utilization of adjacent services could indicate likelihood of using another health service at the same facility.
- Age positively correlated with likelihood to enroll, which is counter to what one might think. As someone ages, they may struggle more to access health services. Conversely however, this may be a factor of the health service itself whereby a younger patient may feel less inclined to follow-up on a chronic disease referral.
- Barriers to cost influenced predictions in the anticipated manner – that facilities where telecounselors noted patients mentioning higher cost and accessibility barriers correlated with referred patients failing to follow through to enrollment.

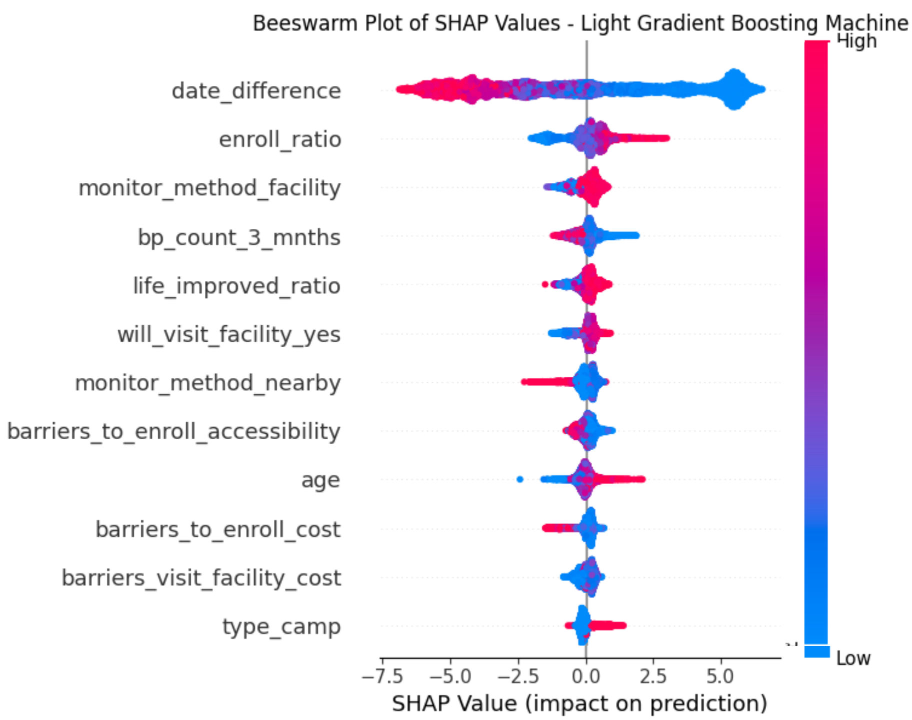


Figure 2. SHAP values for the first 12 features of our model. For each value of a given parameter (indicated by blue-low vs red-high values), its SHAP value is then plotted according to the magnitude of how positively or negatively that value contributed to the patient's enrollment prediction. A feature whose color spectrum that is smoothly distributed insinuates a

positive or negative correlation with the target variable. This is easily demonstrated by the distribution of color in the first feature, date_difference, which negatively trends with enrollment likelihood (the larger the date difference, the less likely the patient will enroll).

2.2 Milestone 5: Product Integration, Interventions & Impact

Process*:*

With the modeling complete, we came up with an initial approach for how this enrollment prediction score could be integrated into our SPICE Technology Suite. At a high level, we want to actionably intervene for patients who do not engage with the health services. An interventionwe actively use is to complete follow-up phone calls for the patients who were referred but have not yet been enrolled. Using our enrollment prediction score, we can then stratify patients in order to decide for whom a call will be most impactful.

Once the production pipeline was ready, we pushed out a release to our SPICE Engage platform to re-define the logic for how patients appear in the list for screening follow-up. To accomplish this, we store the enrollment prediction score in a separate table in the SPICE database. The prediction score is then loaded along with patient details when they get loaded into SPICE Engage, allowing the application to sort the patients for a call based on enrollment likelihood.

Insights:

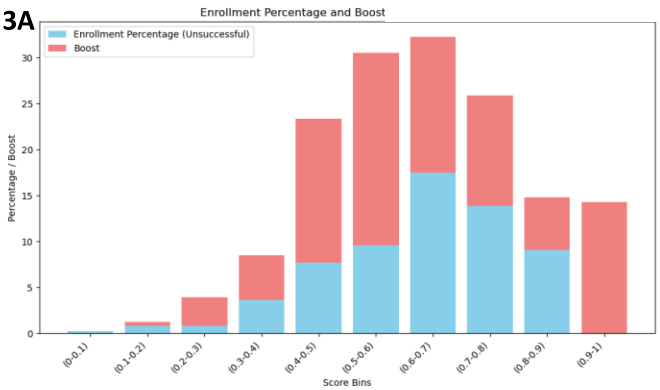
Call Impact with respect to enrollment likelihood: We gained insight for which patients a follow-up phone call would be most impactful with respect to their enrollment prediction score. Productionalizing the model and displaying the enrollment prediction scores within the SPICE database was completed at the end of August. Prior to that date, we also completed analysis of the impact that previous calls had had on patients with respect to their enrollment prediction score to determine how it would inform our sort logic for which patients to call.

From our analysis, we gained the insight that patients with enrollment prediction scores between 40-80% saw the greatest boost to their enrollment numbers after a phone call. This can be concluded from Figure 3.

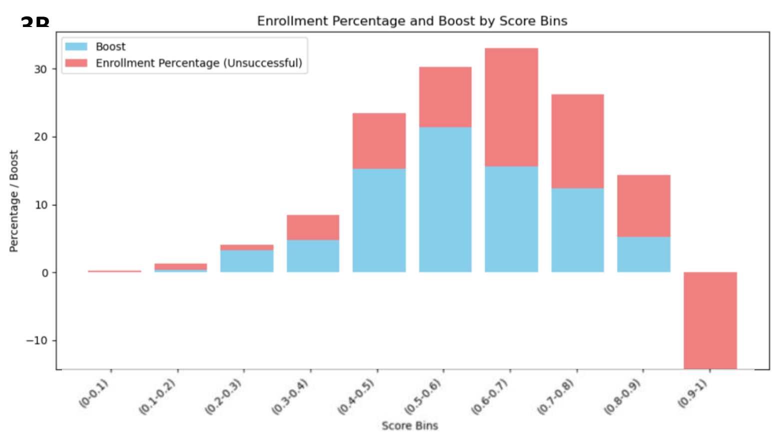
Call Impact with respect to days post screening: We gained insight for which patients a follow-up phone call would be most impactful with respect to their days post screening. We also looked at effectiveness of calls with respect to the days post screening for patients. As can be seen in Figure 3C, patients who were called saw the greatest boost to their enrollment when called within 20 days of their screening.

Outcomes from our calling experiment: We gained insight as to how we can prioritize calls in the future for the greatest impact. From the insights mentioned above around the best time to call and the score buckets most impacted by a call, we implemented these rules into our SPICE Engage product by prioritizing patients in the 40-80% call category less than 20 days since their screening. We applied this only to our telecounselor's view of the platform (we did not adjust which calls the facilities made.) However, once it was calculated how many patients would remain on the platform, we loosened the rules a bit so that the telecounselors would not run out of calls to make. Namely, we extended the window to include higher likelihood of enrollment (since historically very few of these calls had been made), and we extended the call window to 80 days. These patients were sorted in descending order of likelihood to enroll.

3A



3B



3C

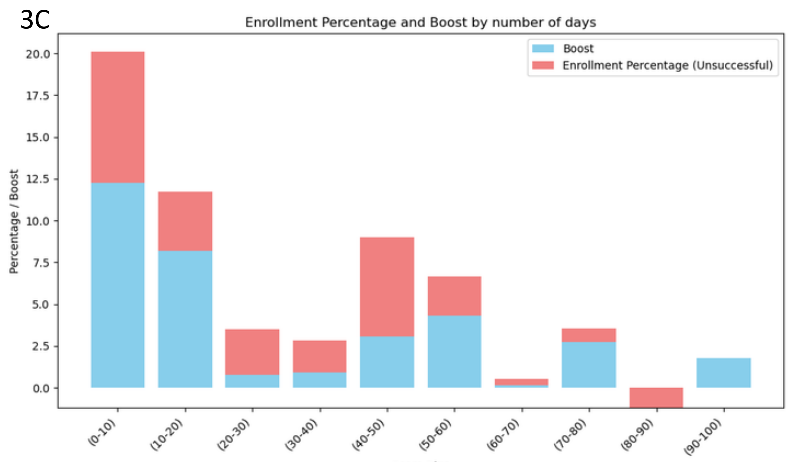


Figure 3. *3*A (top) The blue bars in figure A show enrollment % of patients who never received a successful call, and thus enrolled without any intervention. The top of the red bar shows the enrollment of the patient population that received at least one successful call. The difference between these two bars can be attributed to the effect of the phone call. Figure 3B (middle) shows the "boost" to enrollment stacked at the bottom of the chart, so that the relative effect of calls with respect to a patient's enrollment prediction score. This helps to highlight that the boost to enrollment is greatest for patients between 40-80% likely to enroll. Figure 3C (bottom) shows the boost to enrollment when stratifying by the number of days post screening when the call was made.

The above-described rules were implemented, and patients were called over the course of the month of September until our analysis date on October 26th. As of October 26th when this analysis was done, Telecounselors (LABS employees) made 970 successful screening calls, and users from the facilities made 936 successful screening calls. From these calls, 226 patients came for enrollment which is 11% in total. But if we dive a bit deeper, we can see 29.4% boost to enrollment in the >90% category. While we were expecting the results to be similar to what we saw during our analysis, most of our boost to enrollment came from calling patients >80% likely to enroll. This can be seen from Figure 4.

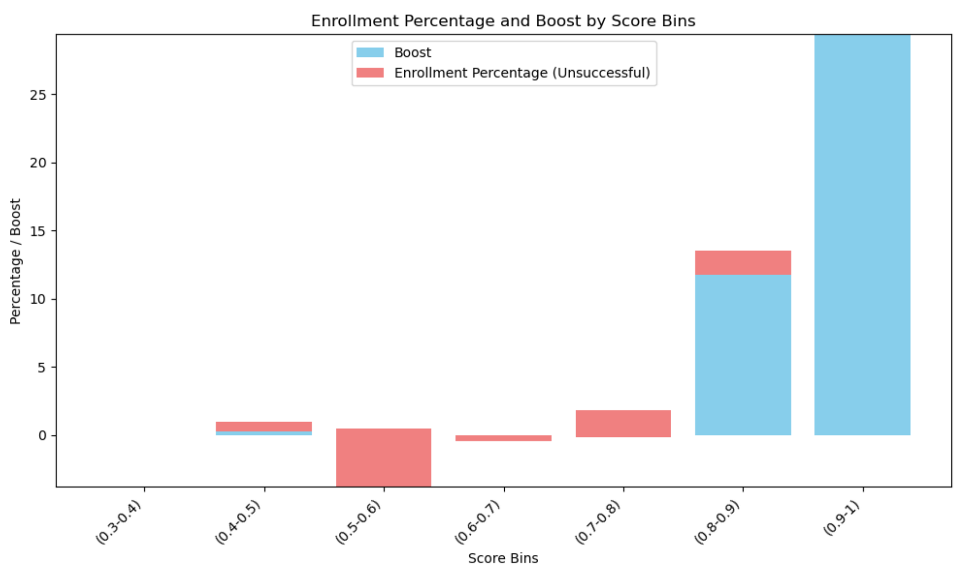


Figure 4. Shows the outcome of our first experiment where we prioritized calls in descending order for their likelihood to enroll. The pink bars represent the patients that came in despite not receiving a successful call, and the blue bar represents, in addition to that, what percentage of patients came in after receiving a successful call.

The insight we gain here is that with the limited number of calls we can make, we can decide on a stopping algorithm in terms of when the efforts give diminishing returns that no longer seem efficient. Table 5 displays the impact on enrollment gained from calling patients with respect to their enrollment likelihood score.

Enrollment Likelihood Score Bucket (0-1)	# enrolled / called	Successfully called enrollment % enrolled/successfully called	Unsuccessfully called enrollment % enrolled /unsuccessfully called	Boost to enrollment from calling
<0.4	28/1010 = 2.7%	S: 9/590 = <1%	U: 19/420 = 4.5%	-3.5%
0.4-0.8	103/2635 = 3.9%	S: 74/1160 = 6.9%	U: 29/1475 = 2%	4.9%
0.8-0.9	23/172 = 12.2%	S: 17/81 = 21%	U: 6/151 = 4%	17%
>0.9	5/29 = 17.2%	S: 5/17 = 29%;	U: 0/12 = 0%	29%

Table 5. This table shows the effect that calling has on enrollment for patients with bucketed enrollment prediction scores. The final column calculates the % of the population that enrolled which can be attributed to the effect of having received a successful call.

Now interpreting these results in Table 5, we can design new heuristics for calling patients for maximal impact. Instead of placing all 1848 calls which yielded 117 enrolled, we could have stopped at:

>0.9 score:
Top 17 calls or 0.9% of the total work with 4.3% of the outcome achieved (x4.74 more efficient)

>0.8 score:
Top 98 calls = 5% of the total work with 18% of the overall outcome achieved (x3.6 more efficient)

>0.7 score:
1101 calls or 60% of total work with 82.5% of the overall outcome achieved (x1.375 more efficient)

>0.4 score:
1258 calls 68% of total work with 92.3% of the overall outcome achieved (x1.35 more efficient)

3.1 Overall Project learnings and final questions

From the beginning, understand the decision you will support: An incredibly helpful learning was the early exercises to understand for whom your model's insights or predictions will have meaning and how do you intend to action that insight. In our case, we understood that knowing who would be less likely to enroll would allow us to intervene. But how, and in what order? Wrestling with that decision of what exactly to do with the uncovered insight of who would not enroll was exceptionally edifying, especially in the presence of other accelerator organizations who were grappling with how to take their insights to the implementation stage. These were some of our favorite conversations during the grant and helped us appreciate the value that on-the-ground knowledge will bring to your model's performance.

Advancing Human Rights Safeguarding through Data Insights: The insights derived from our data exploration and modeling process hold potential for advancing the safeguarding of human rights in healthcare. By unpacking the factors influencing patient enrollment and engagement, we can tailor interventions to address specific challenges faced by individuals in accessing health services. For instance, our identification of the impact of barriers to cost on enrollment predictions suggests a need for targeted interventions to alleviate financial burdens that may impede patients' willingness to engage with healthcare programs. We identified that patients' engagement with health services tangent to the one of direct interest gives an analogous read on patients' engagement with that facility in general. This could help vet likelihood of success of a new health delivery program given success metrics from previous programs. Moreover, our focus on understanding the dynamics of patient demographics, such as age, BMI and the correlation with enrollment likelihood, can inform more inclusive and equitable healthcare strategies. These insights contribute understanding for organizations to address specific challenges faced by patient populations, fostering inclusivity in healthcare access.

Future Impact of Data Approach on Team's Work: The experience gained from this data approach has laid a robust foundation for future data science work within our team and organization. The utilization of advanced modeling techniques and cloud computing services not only allowed us to build our first AI model for integration into our products, but also equipped our team with valuable skills in working together and handling complex datasets efficiently. The awareness of potential biases in the dataset, as evidenced by the imbalance in enrollment days, has heightened our team's sensitivity to exploring and wrangling our data before simply using it for training. Moving forward, this experience will serve as a learning experience, strengthening our ability to navigate and mitigate biases in predictive models, ensuring responsible and fair deployment. Additionally, the successful transferability of our model across different countries underscores the adaptability of our approach, laying the groundwork for scalable and globally applicable solutions. As we integrate enrollment prediction scores into our SPICE Technology Suite, the team gains a powerful tool for targeted interventions, enhancing the impact of our health programs and promoting more effective resource allocation.

Application of Learnings to Similar Nonprofits Facing Data Challenges: The challenges we encountered and insights gained during this project could offer valuable lessons for other nonprofits at a similar stage in their data journey. Our experience highlights the importance of addressing data quality concerns from the outset while emphasizing the power of on-the-ground knowledge and command of the data collection. Our approach of using multiple models to compensate for data gaps could be critical for organizations expanding into new regions or facing changing data landscapes. The use of different models based on data availability and characteristics enables flexibility and adaptability. Moreover, our exploration of explainability through SHAP values provides an example of how to understand the nuanced impact of features' values on a model.