Using Machine Learning to Help Improve Community Health Worker Operations: A Collaboration with M'Care

Joyce Luo JOYCELUO@MIT.EDU
Operations Research
Massachusetts Institute of Technology

Angela Lin AGLIN@MIT.EDU
Operations Research
Massachusetts Institute of Technology

Elizabeth Whittier BETHWH@MIT.EDU
EECS
Massachusetts Institute of Technology

Darcy Kim DARCYKIM@MIT.EDU
Math
Wellesley College

1. Introduction

1.1. Project Background

M'Care provides equitable access to healthcare services such as early detection, diagnosis, and treatment in Nigeria's economically disadvantaged communities. They utilize live AI decision support to help community health extension workers provide personalized healthcare in rural communities (Abramian).

In Nigeria, chronic malnutrition leading to impaired growth affects 32% of children, but only 1 out of 5 children affected are provided treatment. Treatments and interventions for malnutrition under age 5 include disease prevention, specially-formulated nutrient-dense therapeutic foods, and supplemental vitamins and micro-nutrient powders (Lenters et al., 2016). Organizing the administration of interventions will help decrease early childhood malnutrition and the consequential health complications.

M'Care is addressing this issue by running studies involving healthcare workers administering doses of micro-nutrient powder, vitamin A, and zinc to children under five. M'Care would like to leverage the data collected from their current operations to maximize the number of children they can successfully deliver vitamin and nutrient interventions to. Our team was tasked with predicting whether beneficiaries (child-mother pairs) would "complete" their prescribed intervention series, as well as predict the aggregate volume of contacts over time. We used various binary classification methods to predict whether beneficiaries would complete their next scheduled contact. This allows M'Care to target their reminders to beneficiaries that are at risk of not receiving their next scheduled contact. We also used time series models to predict the aggregate volume of contacts, which helps

M'Care to plan ahead the number of resources (vitamins/nutrients, health workers) to have available at any given time.

1.2. Contributions

Our contributions are the following:

- Analyze data from M'Care's current operations to understand current health worker workload and intervention completion rates
- Predict whether each beneficiary will complete their next recommended contact/dose of an intervention
- Predict aggregate volume of expected contacts over time
- Discuss implications and make recommendations to M'Care

2. Related Work

2.1. Binary Classification in Healthcare

Binary classification models such as logistic regression, tree-based methods, and neural networks have often been used in predicting clinical outcomes, such as for predicting the presence of cancer or mortality after surgery, or mortality within 24 hours Liu et al. (2022). Some critical steps in the approach are identifying and selecting variables, assessing model performance, performing internal and external validation, recalibrating the model, and assessing the clinical impact of the model. We train logistic regression, random forest, and neural network models for binary classification to predict whether a patient will receive their next expected contact.

2.2. Time Series Prediction

A time series is a series of measurements of a certain quantity at different points in time and can be decomposed into long-term trends, seasonal/cyclic variations, and random fluctuations or noise. Time series prediction can involve variables other than the target variable itself that changes over time, and the prediction can be single step or multi-step. Some common models for time series prediction include auto-regressive and/or moving average models such as auto-regressive integrated moving average (ARIMA), exponential smoothing models, generalized linear models, and deep learning models. (Lazzeri, 2021). We train auto-regressive models, exponential smoothing models, and a tree-based model (XGBoost) for time series prediction to predict the number of contacts on a given day.

Prediction intervals, which capture the uncertainty around pointwise predictions, can help guarantee for instance, with 95% probability, the true quantity at a particular time will lie within an interval. Sampling different sets of observations from the time series can help provide probabilistic prediction intervals, which is implemented in the Python package `GluonTS` in combination with a deep auto-regressive (DeepAR) model (Alexandrov et al., 2019, 2020).We utilize this package to get prediction intervals.

Finally, ensembling predictions from different models often helps overall predictions by leveraging uncertainty in predictions and combining the strengths of each model. Bertsimas and Boussioux (2023) develops an advanced dynamic robust ensembling method for time

series prediction that dynamically adjusts the weights for the models over different time steps. We utilize both the static and dynamic ensembling methods to ensemble our time series models.

3. Data and Experiment Setup

3.1. Data Description

The data was collected with the M'Care app and covers service delivery between September 2021 and October 2022. The data has been de-identified for privacy and only essential personal details for monitoring dosage and time series have been left. The features are:

- Beneficiary ID = replaces name/identifying information of beneficiary
- Beneficiary type = age group
- DOB = date of birth
- Gender = male or female
- Service name = pharmaceutical intervention (medicine) provided
- Contact number = number of times each beneficiary was visited so far in the data set
- Delivery date = date beneficiary received the service according to the health worker
- Age on delivery date = age (in days) based on delivery date and DOB
- Sync date= date and time (to seconds) that the record was synced to server
- Healthcare Worker ID = ID that represents an individual health worker
- Location ID = village number (geographic area)

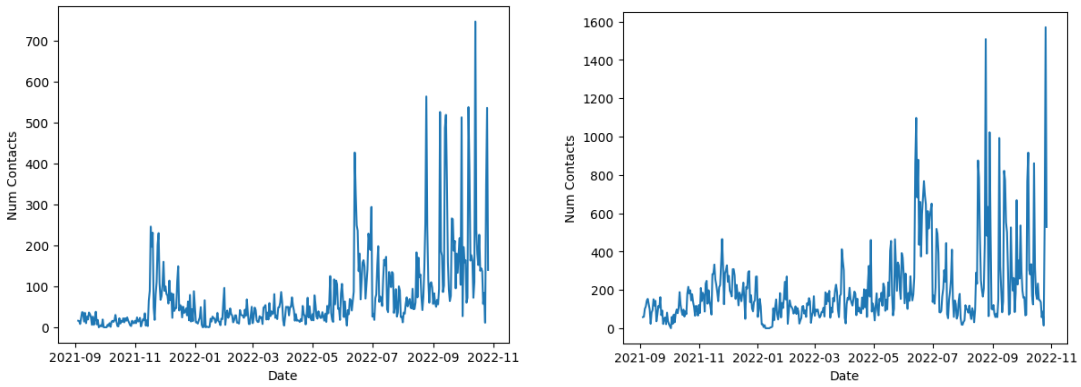
We have data for two different service interventions over the same time period: micronutrient powders for children 6-24 months (MNP) and Vitamin A supplement for children 6-59 months (VITA).

3.2. Exploratory Data Analysis

We have 24K unique beneficiaries for the MNP data set and 58K unique beneficiaries for the VITA data set. We have a total of 56 health workers, each of which exclusively work in 1 of the 8 locations. The locations have 9, 11, 7, 8, 5, 10, 4 and 2 health workers respectively.

The micronutrient intervention is supposed to be given to children from 6-24 months of age every 4 months; however, we see that 83% of patients only received the intervention 1 time, largely due to the fact half of the patients' first dose was at age 25 months. The vitamin A intervention is supposed to be given every 6 months between the ages of 6 months and 5 years. We see in our data that about half of patients receive the intervention once and half receive it twice. In the Appendix, we show the distribution in age at first contact (first time the intervention was given).

We aggregated the total number of contacts/observations per day across all locations and health workers to create an aggregate time series from 9/04/2021 to 10/27/2022. We can see from Figure 1(a)subfigure and Figure 1(b)subfigure that there is no smooth trend to the number of contacts per day. The number of contacts per day seems to vary drastically, and there are jumps from the low hundreds to the thousands, especially for the VITA intervention.



(*a*) MNP (*b*) VITA

Figure 1: Total number of contacts per day from 9/04/2021 to 10/27/2022.

We also plotted the daily time series of number of contacts for each location and for each health worker. The plots over time for each location have similar shapes but different heights (volume of number of contacts). The uneven distribution of contacts across different health workers could indicate an opportunity/need for better planning/work assignment across the health workers. In the appendix, we show the time series for the 5 health workers in location 5.

We were able to provide analysis on the number of contacts performed by each health worker in each month to M'Care which helped them understand the variation in performance across locations and individual health workers.

4. Methods

We have two approaches for using machine learning to help M'Care improve their operations. One of these approaches focuses on prediction at the individual beneficiary level (where we consider a single beneficiary as a child-mother pair), and the other approach focuses on time series prediction at the aggregate level.

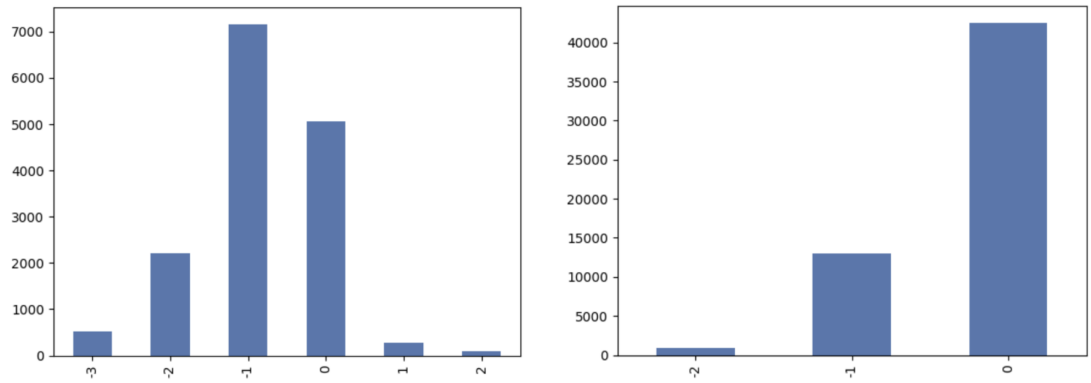
4.1. Predicting Beneficiary Follow-up Rates

Given a variety of beneficiary-level features (described in Section 3.1), we aim to predict whether the beneficiary will show up to their follow-up contact. To calculate our label we compare the expected number of contacts based on dosage schedule with the actual number of contacts made. Expected number of contacts is calculated based on the date of first contact with the patient and the date the patient ages out of the program or the last date we have collected data from if they do not age out. If the patient's next contact is expected to be after the last date we have collected data from, the data of the patients last contact is excluded from the training set. The features we use are the beneficiary's region, time of previous contact, patients age, and gender, relevant holidays, and if the contact is earlier or later than expected. Many of the features we use for classification required some feature engineering, such as the calculation of expected number of contacts, and type I censoring.

Only using data for patients who are expected to have more than one contact, we split the data into train, test, and validation sets.

4.1.1. Completion Rate

When calculating completion rates, we excluded any participants whose expected number of contacts = 1 (due to either beginning the study at the end of the data collection period or aging out after receiving the first dose), as they are irrelevant to predicting patient follow up. Completion has been defined in a binary way, where either the participant meets the expected number of contacts or does not, but Figure 2 illustrates a participant's deviation from the expected number of contacts to see the degree of partial completions.



(*a*) MNP (*b*) VITA

Figure 2: Deviation from Expected Number of Contacts.

Expected Number Contacts	MNP	VITA
2	32.29%	94.15%
3	48.51%	14.45%
4	17.15%	—
Total	35.47%	75.41%

Table 1: Completion Rates

4.1.2. Binary Classification

Helping M'Care predict which beneficiaries will show up for further contacts will allow them to target resources towards beneficiaries that are less likely to return for follow-ups. For both the MNP and VITA datasets, we trained three binary classification models that determine if a beneficiary will return for a follow up: logistic regression, neural network (NN), and random forest classifier (RFC). We used an 80-20 train-test split using beneficiary ids to ensure there is no data leakage. The neural network consists of two fully connected RELU layers. Although neural networks are not explainable, we can interpret feature importance from both logistic regression and random forest classification models. For logistic regression,

feature importance is determined by the magnitude of the coefficient of each feature. For random forest, feature importance is determined by mean decrease in impurity. For the RFC, we also ran a hyper-parameter search on the number of decision trees in the forest and the maximum depth of each decision tree in the forest.

4.2. Predicting Aggregate Number of Contacts

We use time series models to predict the daily number of total contacts (across all health workers) for each intervention. We aim to provide to M'Care a useful model that will help them predict the total/aggregate amount of contacts their healthcare workers are carrying out and what the variation in this could be so they can adequately plan their resources.

We learned from our mentors at M'Care that they onboarded many new health workers during the time period our data was collected and therefore the variation in number of total contacts could be largely due to the number of health workers at any given time. Therefore, we decided to normalize our time series by the number of health workers working at each time, so our new time series values are the average number of contacts per health worker on each day.

4.2.1. Additional Features for Time Series Prediction

The standard features used in time series prediction are just the values of the time series in previous periods, i.e. the lags. However, we also are able to add additional features to models in addition to just past observations. We consider the binary variable representing whether or not the day is a weekday (i.e., 1 if it is a weekday, and 0 otherwise). After discussion with our M'Care collaborators, we were also able to add features that were more specific and relevant to our data set. For instance, our M'Care collaborators mentioned that healthcare workers would not be working on Nigerian holidays, so we created a holiday indicator variable. They also mentioned that they run a promotion every month at the end of the month to encourage the healthcare workers to get more contacts, so we created an indicator for the last 3 days of the month. Additionally, they mentioned that on Thursdays and Fridays healthcare workers tended to input more contacts, so we created an indicator for Thursdays and Fridays as well.

4.2.2. Time Series Models

We use autoregressive models with and without additional features, Holt-Winters exponential smoothing, and XGBoost as our first models. An autoregressive model (AR(p)) is a linear model where the predicted value in the current period is a function of the values of p past time periods, also known as the "lag". This is the simplest possible time series model, and we chose a lag of 15 for the MNP intervention and a lag of 10 for the VITA intervention due to the fact that the autocorrelation is significant up to around 15 lags for MNP and up to around 10 lags for VITA (figure in Appendix).

We also use Holt-Winters exponential smoothing (HWES) as another baseline model. HWES considers 3 aspects of a time series: levels, trends, and seasonality. This corresponds to the 3 types of HWES: Single, Double, and Triple. Triple HWES is able to model a time series with levels, trends, and seasonality, Double HWES can model levels and trends, and finally Single can only model a time series with levels (Chatfield, 1978). After analyzing the

decomposition of our time series into levels (figure in Appendix), trends, and seasonality, we chose to use Triple HWES with additive seasonality, additive trend, and 50 seasonal periods.

For XGBoost, we split the data in an appropriate training and test set with 15 lags as 15 different features and additional features we created. We set the number of estimators as 100, maximum depth as 5, and the learning rate at 0.1. Since XGBoost is a one-step prediction algorithm, we use the value in the previous time period as the initialization.

We also ensembled the predictions from our 4 previous models, i.e. we took a weighted average (convex combination) of the predictions by each model to get our final prediction for each time. We build a static ensemble by tuning the (static) weights for each model on a validation set. We also implement a dynamic ensembling method from Bertsimas and Boussioux (2023), where the weights of the ensemble are fitted to dynamically adjust over time.

4.2.3. Prediction Intervals

The models described above only give pointwise predictions, but we also aim to get prediction intervals around the pointwise predictions. We use the DeepAR model in combination with the `GluonTS` package in Python, which uses recurrent neural networks (RNNs) and provides probabilistic prediction intervals (Alexandrov et al., 2019, 2020; Salinas et al., 2019). We set the number of samples to 100, and we test this model for both 50% and 90% prediction intervals with the additional weekday feature.

4.2.4. Evaluation

We split the data into training and testing sets by treating everything but the last 30 days as the training set and the last 30 days as the testing set. We use weighted mean absolute percentage error (wMAPE) as the evaluation metric for all of our models. wMAPE is described by the following formula, where $\mathbf{A}$ is a vector of the true data and $\mathbf{F}$ is a vector of the model forecasts: $\text{wMAPE} = \frac{\sum_{i=1}^{T} |A_i - F_i|}{\sum_{i=1}^{T} |A_i|}$. Our data analysis showed that the data we use as the test set looks drastically different than the rest of the time series. Therefore, we also tried shifting the training and testing windows to see if model performance was drastically affected by this. We additionally trained our models using only the first 200 days of data and then predicted the next 30 days.

5. Results

5.1. Binary Classification Results

The RFC is the best model with a 0.99 AUC for both MNP and VITA.

5.1.1. MNP

Table 2 summarizes model performance. For both logistic regression and RFC, Contact Number was by far the most important feature in prediction. Further feature importance breakdown for logistic regression can be found in the appendix. As evidenced by Table 2,

the RFC is the best model for predicting MNP follow up rates. Figure 3(a)subfigure further elucidates feature importance in the RFC model.

5.1.2. VITA

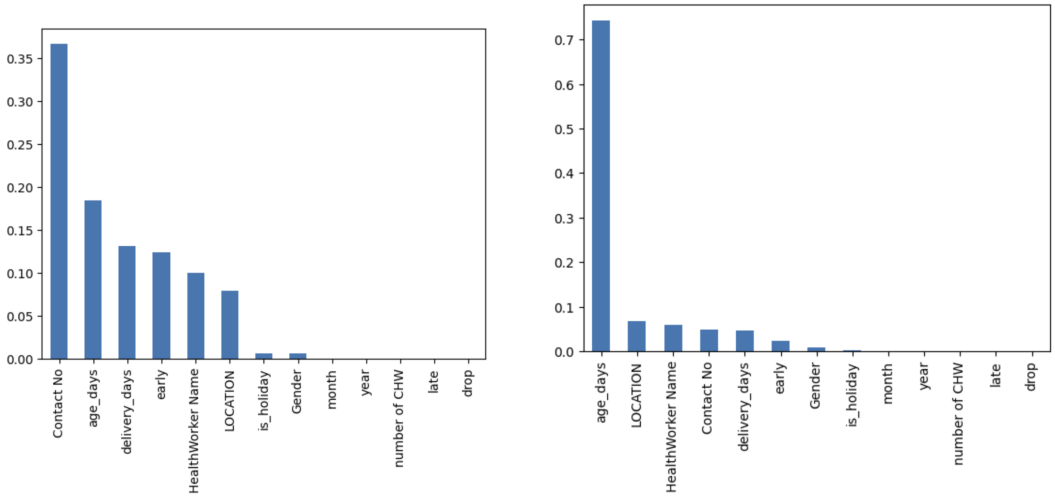
Table 3 summarizes model performance. For the logistic regression model, Contact Number was again the most important feature in prediction. Further feature importance can be found in a figure in the appendix. Figure 3(b)subfigure further elucidates feature importance in the RFC model. For this model, age is by far the most important figure. The older a patient is, the less likely they are to follow up.

Model	tn	fp	fn	tp	Recall	Precision	f1	Accuracy	AUC
Log Reg	64%	2.8%	0.51%	33%	0.98	0.92	0.95	0.97	0.99
NN	56%	11%	6.6%	27%	0.80	0.72	0.76	0.83	0.92
RFC	65%	1.8%	0.51%	33%	0.98	0.95	0.97	0.98	0.99

Table 2: MNP Model Performance

Model	tn	fp	fn	tp	Recall	Precision	f1	Accuracy	AUC
Log Reg	5.7%	0.28%	0.56%	93%	0.99	0.99	0.99	0.99	0.99
NN	3.8%	2.2%	0.7%	93%	0.99	0.97	0.98	0.97	0.97
RFC	13%	1.4%	1.1%	85%	0.99	0.98	0.99	0.97	0.99

Table 3: VITA Model Performance



(a) MNP (b) VITA

Figure 3: Feature importance from random forest classifiers.

5.2. Time Series Results

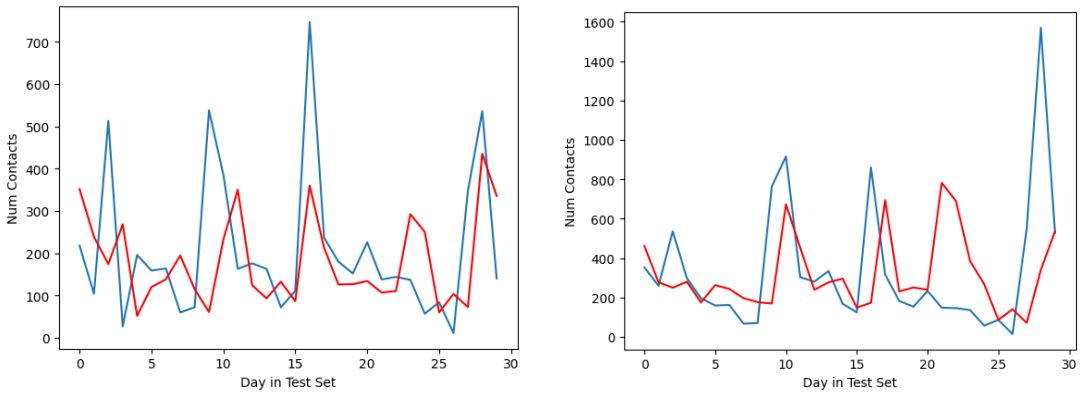
5.2.1. Baseline Results

As an initial baseline comparison, we tested out simpler time series prediction models to get an idea of how well we can predict the total number of contacts per day for each intervention. We first tested out the AR(p) model for both interventions. Table 4 shows the wMAPE performance metric from testing the AR(15) model for MNP and AR(10) for VITA. We can see that our preliminary simple models do not do well on the testing data. An acceptable value of wMAPE should be around 0.4 or below.

Model	MNP	VITA
AR(p)	0.5721	0.6292
AR(p) + Weekday Feature	0.5490	0.6394
Triple HWES	0.5334	0.7241
XGBoost + Weekday Feature	0.6298	0.6827

Table 4: wMAPE prediction metrics for different baseline models.

We can also see that performance is worse for the VITA intervention dataset compared to the MNP dataset for all 3 baseline methods that we tried. Comparing Figure 4(a)subfigure to Figure 4(b)subfigure, we can see that the jumps in the VITA data are larger than the MNP data, which is likely why model performance is better for the MNP data. From Table 4, we can see that the model performance improves slightly for the MNP intervention when adding the weekday feature to the AR(p) model, but actually gets worse for the VITA intervention. We also test Triple HWES, and we can see that once again, model performance improves for the MNP intervention but gets worse for the VITA intervention (as shown in Table 4). For both the AR(p) model and Triple HWES, the predictions are very conservative in comparison to the actual data values (figure in Appendix).



(a) MNP (b) VITA

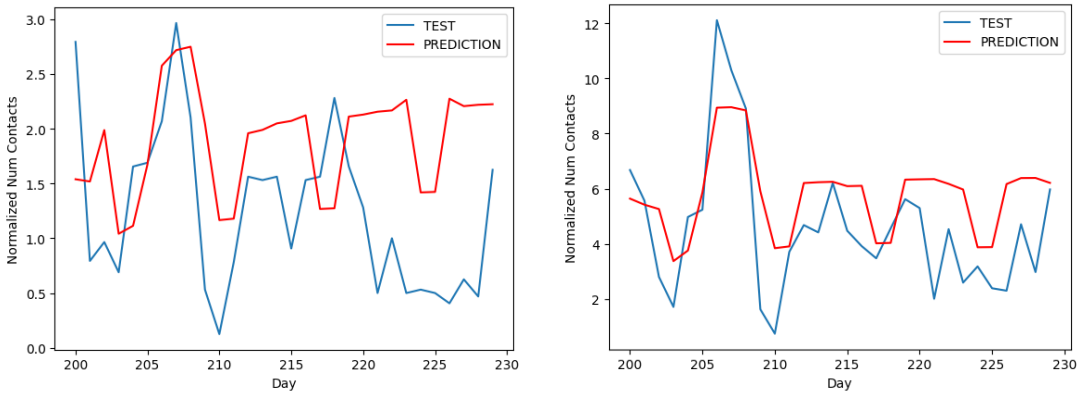
Figure 4: XGBoost predictions compared to the test data.

The final baseline model we tested was XGBoost, and we can see that its performance in terms of wMAPE seems to be worse than the other models. However, Figure 4 shows that the model predictions seem to follow the trend a bit more than the other models and captures the magnitude of the peaks. This is likely because XGBoost is just using the previous value as the prediction for the next value, which substantiates why the wMAPE is worse even though the plots look better.

5.2.2. Modified Models Results

We now show our prediction results with the normalized data (predicting average number of contacts per health worker), additional features, and adjusting our training/prediction window to the first 230 days. For the AR models, we used $p = 8$ for MNP and $p = 7$ for VITA. Comparing Table 4 and Table 5, we can see that notably, the performance of the models on the VITA dataset improves significantly for the AR models and XGBoost. Normalization on its own did not have much of an effect on the performance metrics (based on additional experiments), but it seemed that the additional features and shifting the training window did have an effect on the performance. Triple HWES performs about the same for VITA and worse for MNP likely due to the fact that we cannot include any additional features for Triple HWES.

The dynamic ensembling method did not improve our wMAPE. However, a static ensemble of our 4 models (weights tuned on a validation set) gave an improvement in wMAPE beyond what our best individual model was able to give us. For MNP, the final ensemble gave 0.1 weight to AR(p), 0.1 weight to AR(p) + All Features, and 0.8 weight to XGBoost + All Features. For VITA, the final ensemble gave 0.25 weight to AR(p), 0.5 weight to AR(p) + All Features, and 0.25 weight to XGBoost + All Features. Neither ensembles utilized the Triple HWES model predictions.



(*a*) MNP (*b*) VITA

Figure 5: AR(p) + All Features model predictions compared to the test data.

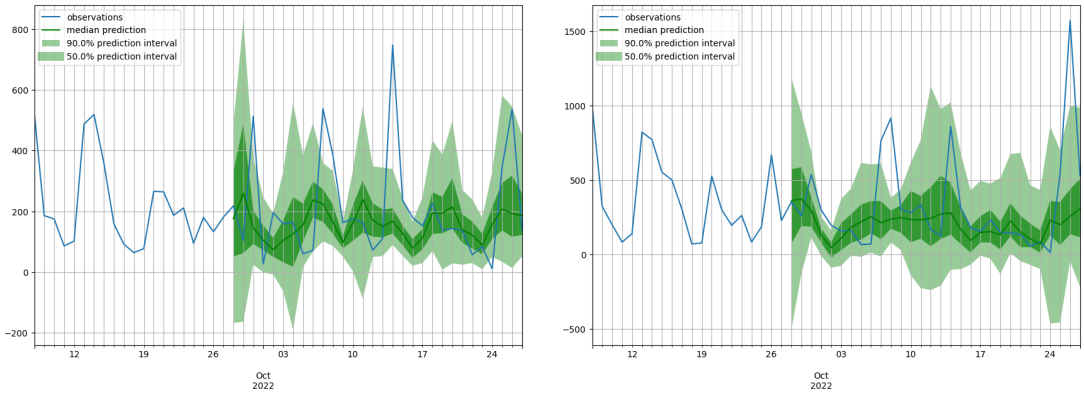
5.2.3. Prediction Intervals

In order to capture inevitable uncertainty in the predictions, we use the DeepAR probabilistic prediction model to obtain prediction intervals. We can see that prediction intervals

Model	MNP	VITA
AR(p)	0.5085	0.3746
AR(p) + All Features	0.7035	0.3643
Triple HWES	0.6722	0.7073
XGBoost + All Features	0.4847	0.3798
Dynamic ensemble	0.7027	0.5014
Static ensemble	0.4807	0.3313

Table 5: wMAPE prediction metrics for models trained on normalized data, with additional features and shifted prediction window.

would be very useful for our current data in order to capture the extreme peaks in the data. Figure 6 shows results from training on the entire dataset except the last 30 days and the daily total number of contacts. The upper bound of the 90% prediction interval seems to capture most of the large peaks in the data, and lower bound of the 50% prediction interval captures some of the troughs in the data. This could be useful for providing rough resource bounds.



(a) MNP (b) VITA

Figure 6: DeepAR model predictions and prediction intervals compared to the test data.

We also test the DeepAR model trained on the normalized data with a shifted prediction window. We obtain an average wMAPE (over 100 samples) of 0.6964 for MNP and 0.5568 for VITA, which is better than the original DeepAR model trained on the entire dataset except the last 30 days. Nevertheless, we noticed from model testing that the prediction intervals did not really capture the peaks for the modified data (figure in Appendix). This result makes sense however, because the original training data (entire dataset except the last 30 days) has a higher variance since it includes data from closer to the end of the time series where there are extremely high peaks. Through sampling, the DeepAR model is able to capture this variance, and therefore provide better prediction interval coverage.

6. Discussion

Insights from Beneficiary-Level Prediction. The original assumption that low completion rates are due to losing patients from studies is contradicted by the completion rates for each contact number. As expected, those with 4 expected contacts have the lowest completion rate, however patients with 3 expected contacts have a higher completion rate than those with 2. We further encourage using the RFC on new and existing patients to determine the likelihood that they will follow up. Identifying patients at risk of not following up before community health workers lose track of them will allow for intervention.

Insights from Time Series Prediction. Our overall results show that predicting the aggregate number of contacts and the average number of contacts per health worker over time is a challenging but feasible task that requires taking into account time series uncertainty. Providing prediction intervals and ensembling provides more robust predictions, since point-wise predictions could have significant error. We illustrate that for this particular data, better point-wise prediction performance comes from considering smaller chunks of more homogeneous data for training and testing windows. However, to get better prediction intervals, the entire time series should be used to capture the variation in observations over time to get more conservative bounds on the number of contacts over time.

Limitations. The main limitations of our study are data-related. We were unable to obtain data related to the beneficiaries' mothers as well as the delivery method (in-home or central location), which could have improved predictions at the beneficiary level. In addition, it would have been useful to have a longer time span of data for time series prediction to capture seasonality throughout the year. The time series trend changed significantly around June 2022 and became much more erratic, which affected our models' performance as shown. Seeing how this trend evolves further in the study would be extremely useful to improve prediction. We also note that we were told later on that the "Delivery Date" field in the data is not the actual date of delivery, but rather the day the contact was inputted into the app by the health worker. We assume that the input date is on or close to the true delivery date in our modeling.

Implications & Recommendations. We provide models both on the beneficiary side and healthcare worker side that could help improve M'Care's operations. Our work could allow M'Care to target their reminders to beneficiaries that are at risk of not receiving their scheduled doses and also help M'Care prepare ahead of time how many resources are needed at the aggregate level at a given future time.

In order to improve the M'Care application and future data collection, we recommend that the data entry fields be expanded to have healthcare workers input the actual date of delivery and method of delivery every time they record a beneficiary contact. We also noticed that while some patients received treatment earlier than expected, none received treatment late (more than 2 weeks after the expected date). This indicates that patients who miss a contact are lost from the study, so M'Care should look into new methods to follow up on missed contacts.

Member Contributions

Joyce and Angela worked on the part of the project related to predicting the aggregate number of contacts (time series). Elizabeth and Darcy worked on the part of the project related to predicting beneficiary follow-up rates (binary classification). More specific contributions of each member are as follows:

- Joyce Luo: cleaning and processing the data for time series analysis; creating additional features for time series prediction; baseline and modified time series models training, prediction, and evaluation; creating time series prediction intervals; framing problem and defining assumptions; data and model interpretations and additional recommendations to M'Care
- Angela Lin: exploratory data analysis on beneficiaries and health workers; creating time series for individual health workers and locations; creating normalized time series from calculating health worker counts; dynamic and static ensembling of time series models; framing problem and defining assumptions; data and model interpretations and additional recommendations to M'Care
- Elizabeth Whittier: cleaning and processing the data for binary classification, creating additional features for binary classification prediction, feature engineering and data censoring for binary classification, completion rate statistics, framing problem and defining assumptions
- Darcy Kim: mapping cleaned data to be compatible for classification models; creating and training logistic regressions, random forests, and neural networks for binary classification; subsequent model analysis and interpretations in context of problem

Acknowledgements

We are extremely grateful for the help we received from our project mentor, Siddharth Srivastava, and M'Care founder, Opeoluwa Ashimi. Their willingness to provide as much information as possible about M'Care's operations and challenges was essential to pushing our project forward. We are also thankful to Noel Shaskan at SOLVE MIT for connecting us to Ope and Siddarth, as well as all of the course staff for their assistance throughout the semester.

References

Jackie Abramian. M'Care Brings Equitable Healthcare Access To Rural Africa With Mobile Connectivity. *Forbes*. URL forbes.com/sites/jackieabramian/2021/02/11/mcare-brings-equitable-healthcare-access-to-rural-africa-with-mobile-connectivity.

A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. C. Türkmen, and Y. Wang. GluonTS: Probabilistic Time Series Modeling in Python. *arXiv preprint arXiv:1906.05264*, 2019.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL http://jmlr.org/papers/v21/19-820.html.

Dimitris Bertsimas and Leonard Boussioux. Ensemble modeling for time series forecasting: an adaptive robust optimization approach, 2023.

C. Chatfield. The holt-winters forecasting procedure. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3):264–279, 1978. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2347162.

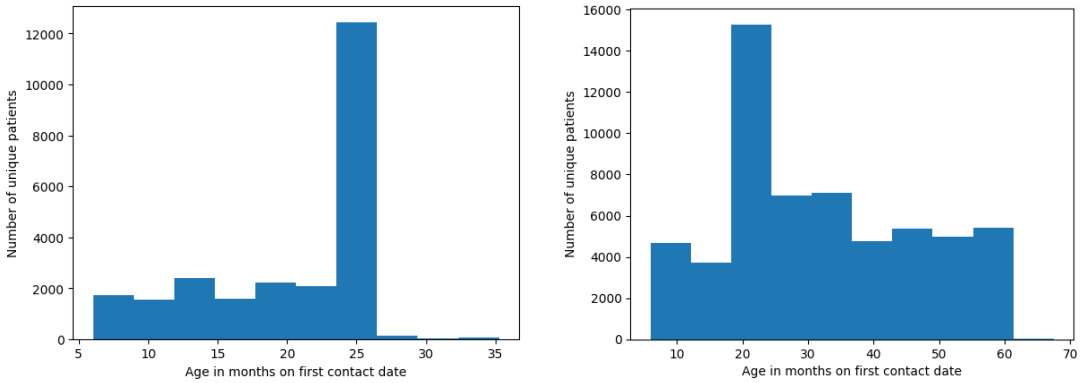
Francesca Lazzeri. *Machine Learning for Time Series Forecasting with Python*. John Wiley & Sons, Inc., 2021.

L Lenters, K Wazny, and Z. A. Bhutta. *Reproductive, Maternal, Newborn, and Child Health: Disease Control Priorities*, volume 2. 3 edition, 2016.

D. Liu, WY. Shin, and E. Sprecher. Machine learning approaches to predicting no-shows in pediatric medical appointment. *npj Digit. Med.*, 2022.

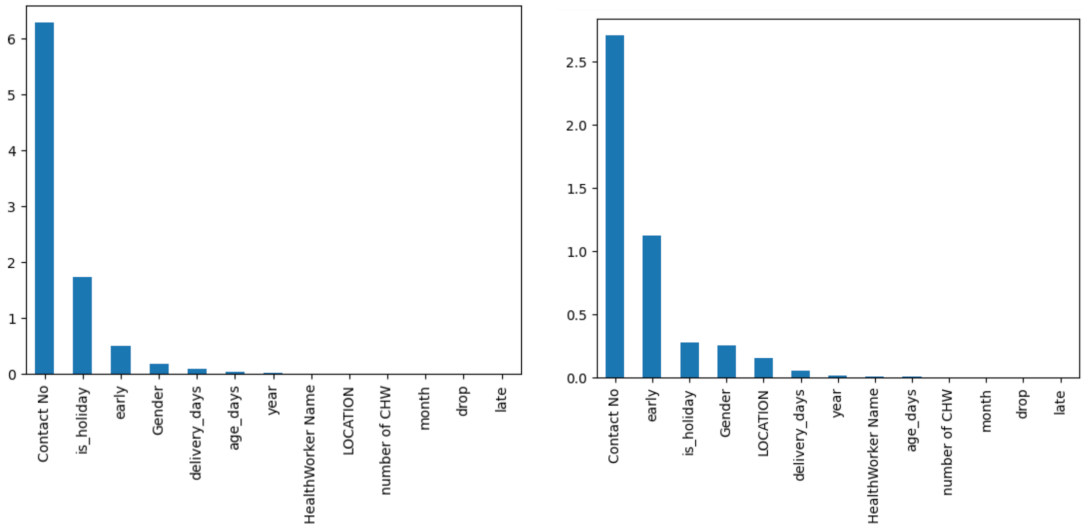
David Salinas, Valentin Flunkert, and Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, 2019.

Appendix



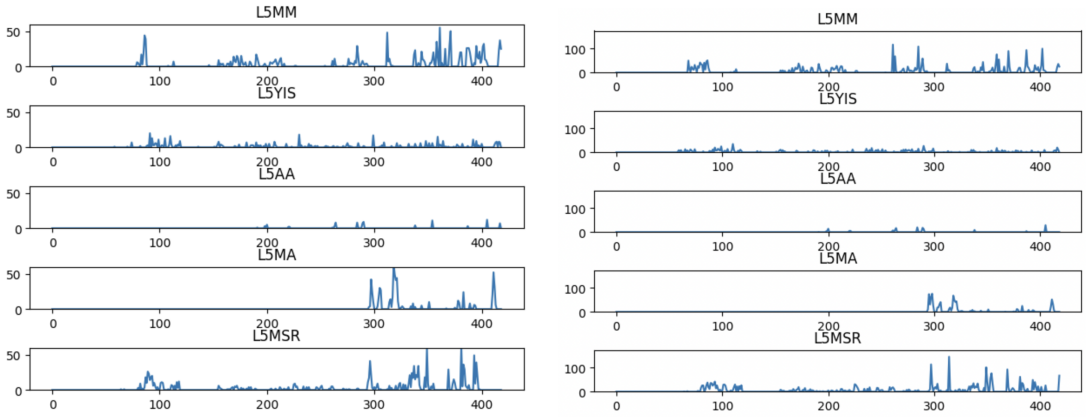
(*a*) MNP (*b*) VITA

Figure 7: Age in months on first contact date



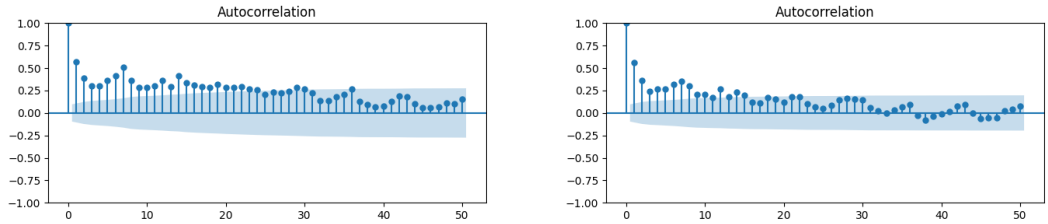
(*a*) MNP (*b*) VITA

Figure 8: Feature importance for logistic regression models.



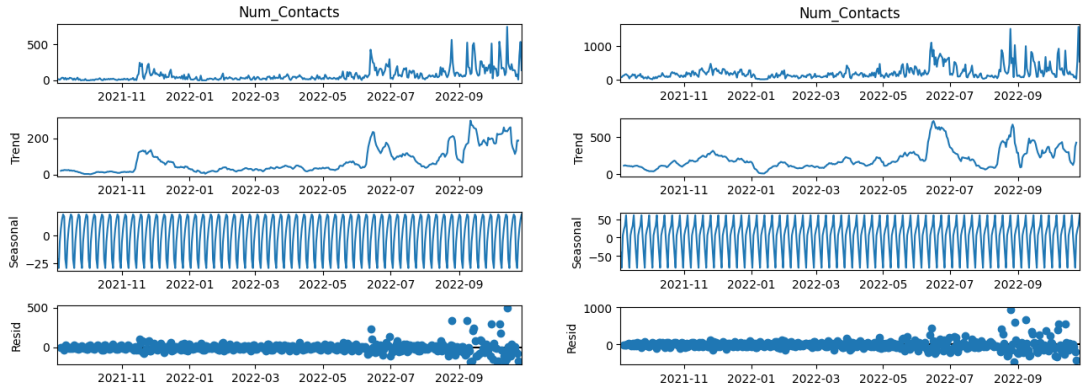
(a) MNP (b) VITA

Figure 9: Time series for each healthcare worker in location 5



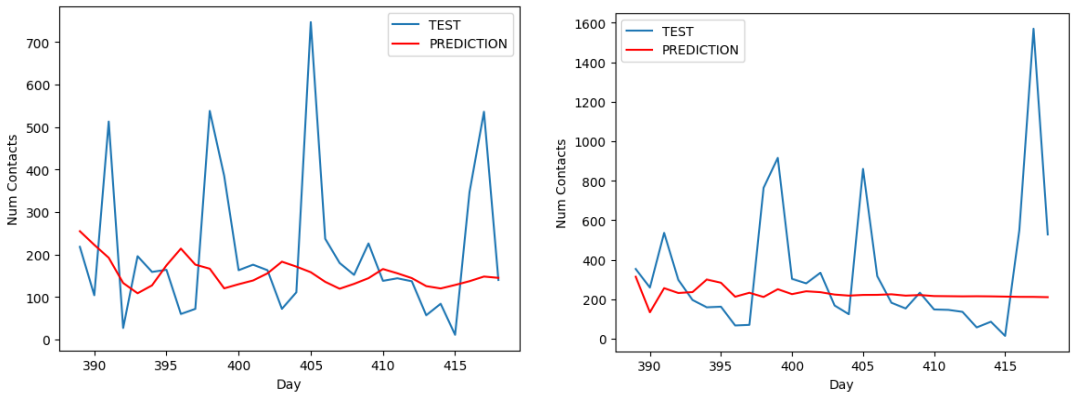
(a) MNP (b) VITA

Figure 10: Autocorrelation plots for both interventions.



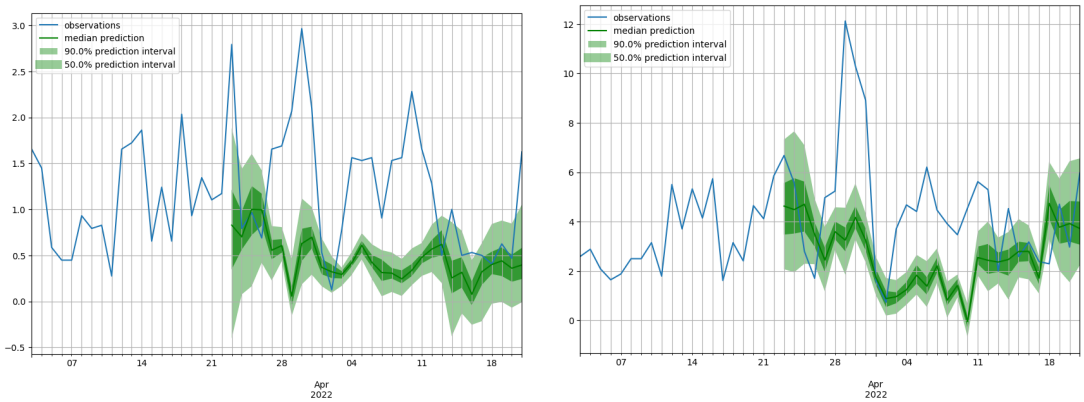
(a) MNP (b) VITA

Figure 11: Decomposition of time series for both interventions.



(*a*) MNP (*b*) VITA

Figure 12: AR(p) model predictions compared to the test data.



(*a*) MNP (*b*) VITA

Figure 13: Modified DeepAR model predictions and prediction intervals compared to the test data.