Predicting Treatment Adherence of Tuberculosis Patients at Scale

Abstract

Tuberculosis (TB), an infectious bacterial disease, is a significant cause of death, especially in low-income countries, with an estimated ten million new cases reported globally in 2020. While TB is treatable, non-adherence to the medication regimen is a significant cause of morbidity and mortality. Thus, proactively identifying patients at risk of dropping off their medication regimen enables corrective measures to mitigate adverse outcomes. Using a proxy measure of extreme non-adherence and a dataset of nearly 700,000 patients from four states in India, we formulate and solve the machine learning (ML) problem of early prediction of non-adherence based on a custom rank-based metric. We train ML models and evaluate against baselines, achieving a $\sim$ 100% lift over rule-based baselines and $\sim$ 214% over a random classifier, taking into account country-wide large-scale future deployment. We deal with various issues in the process, including data quality, high-cardinality categorical data, low target prevalence, distribution shift, variation across cohorts, algorithmic fairness, and the need for robustness and explainability. Our findings indicate that risk stratification of non-adherent patients is a viable, deployable-at-scale ML solution.

1. Introduction and Problem Statement

Tuberculosis is one of the world's great scourges; it was the cause of the second-highest number of deaths from a single infectious agent in 2020, with an estimated 10 million cases worldwide, and 1.5 million deaths (WHO, 2021). In India, with a median estimated incidence of 2.59 million TB cases in 2020, TB is a serious public health concern (WHO, 2021).

While drug-sensitive TB (DSTB) is treatable through a standard drug combination typically administered for six months, drug-resistant TB (DRTB) is much more difficult to treat. A major cause of DRTB is lack of adherence to the standard treatment (Volmink and Garner, 2007; Jain and Dixit, 2008; Mekonnen and Azagew, 2018). In order to drive patient adherence to the treatment regimen, Directly Observed Treatment Short-course (DOTS), or direct observation of the patient, has so far been the mainstay recommendation for successful TB treatment outcomes (MoHFW, 2022).

Non-adherence to TB treatment exists in many forms; an extreme form is loss to follow-up (LFU), representing a treatment interruption of thirty consecutive days or more. In 2020, India's national TB elimination program (NTEP) reported that $\sim$ 3% of TB patients who started treatment became LFU, with upto 13% LFU rates in DRTB patients (MoHFW, 2022). LFU patients pose serious challenges as silent transmitters of TB and contribute to increased risk of morbidity, mortality, and cost burden (Munro et al., 2007; Osman et al., 2021; Zheng et al., 2020).

While rigorous DOTS monitoring and periodic patient home visits are core mandates for TB field staff in India (MoHFW, 2022), providing the same level of intensive

monitoring to all patients is infeasible in the resource-constrained TB healthcare setting, leading to LFU. Other factors causing high LFU include: alcoholism, smoking addictions, previous history of TB, adverse drug reactions, chronicity of treatment and pill burden, lack of family/social support, migrant status, out-of-pocket expenditure, stigma, and poor knowledge of the disease (Jaggarajamma et al., 2007; Washington et al., 2020).

Problem Statement. We model the task of predicting whether a patient p will be LFU as a binary classification problem with target label y_p (1 for LFU and 0 for non-LFU), where patient-level features are represented as $\mathbf{x}_p = \{x_{p1}, x_{p2}, \ldots, x_{pd}\}$. The model outputs a risk score $\hat{y}_p = f(\mathbf{x_p})$, corresponding to the propensity for a patient becoming LFU, where $f : \mathbb{R}^d \rightarrow [0, 1]$. Features of n patients can be collected into a $n \times d$ feature matrix $\mathbf{X}$. True and predicted label vectors are denoted as $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively.

Our solution is a machine learning (ML) tool that recommends early stratification of TB patients at risk of LFU to enable effective differentiated interventions and care, through modified patient monitoring workflows and operating procedures. Risk stratification leads to optimal use of existing resources and enables systematic prioritization of patients for intensified care. Figure 1 shows how the solution fits into the journey of a patient registered in Nikshay, a web-enabled patient management system for TB control run by India's National Tuberculosis Elimination Program (NTEP).

Contributions. The major contributions of our work are as follows:

1. *Problem formulation*: We identify a large-scale public-health problem affecting millions of patients worldwide, and formulate and solve it as an ML problem for benefit allocation in a resource-constrained setting. Deployment considerations and real-world constraints are incorporated into the model to maximize the solution's effectiveness. These include a custom rank-based sensitivity metric for evaluation and model selection, and splitting the data temporally (forward splitting) to mimic the deployment scenario directly.

2. *Model enhancements*: The dataset has large size, low target prevalence, a substantial number of high-cardinality categorical features, and is subject to data distribution shifts due to seasonal variations and a diverted focus of the healthcare system on other epidemics. We extensively search for the best encoding-model pair in this scenario, including state-of-the-art categorical variable encodings, tabular deep learning models, and fully interpretable boosting algorithms. Other techniques that we experiment with include data augmentation, ensembling, interpretability, and algorithmic fairness, enabling us to guide deployment and convince public health authorities.

3. *Evaluation and generalization capability*: We set aside six months of data for a passive evaluation so as to incorporate most out-of-distribution cases. We comprehensively evaluate models across multiple cohorts of interest – geographical, gender-based, type of TB, and public/private health facility. Our models are found to generalize well across time and geographies and provide good measures of predictive multiplicity (Marx et al., 2020) as evidence of the robustness of our model scores.

4. *Potential social impact*: The goal of the project is to systematically deploy the solution across all 780 districts in India, impacting over a million TB patients annually. At each geographical cohort, health workers will monitor a dashboard that displays the risk status for each patient, which field staff will use to provide prioritized interventions. This will allow targeted utilization of health work-

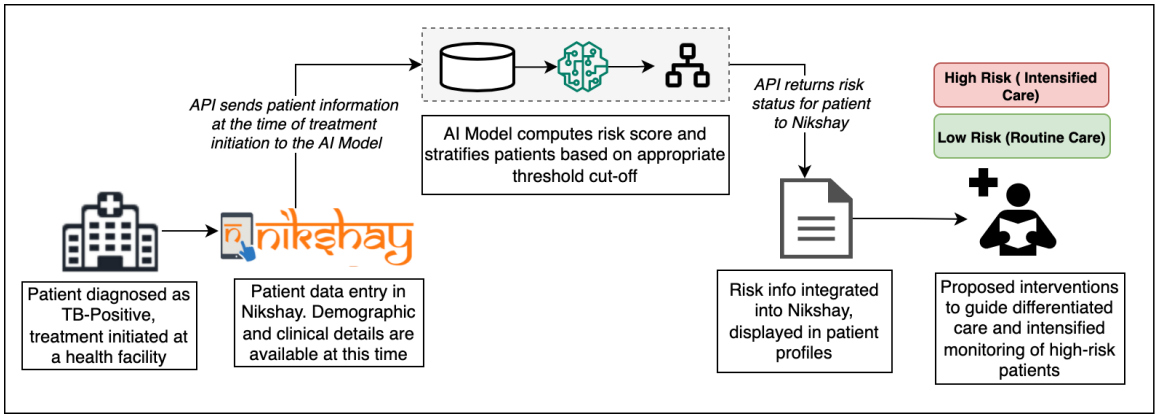


Figure 1: Proposed patient journey, from being diagnosed TB positive to receiving differentiated care based on the risk score computed by our solution.

ers' bandwidth. For 325,190 patients with treatment initiated in the last six months of 2020, our ML solution leads to a potential impact on 5587 lives, compared to 1808 with random targeting and 2974 with the best rule-based baseline model (Section 5.4). We are now working closely with the Central TB Division, Govt. of India, and several Indian states for a pilot deployment.

2. Related Work

There has been significant work on predicting adherence to medication regimens, for HIV/AIDS (Ettenhofer et al., 2010), heart failure (Davis et al., 2014), and mental disorders (Adams and Scott, 2000). For TB, statistical models such as survival analysis or regression have been used across the developing world (Shargie and Lindtjørn, 2007; Kliiman and Altraja, 2010; Roy et al., 2015; Xu et al., 2017). In the Indian context, (Killian et al., 2019) construct ML models to predict TB medication adherence patterns using past data recorded through the 99DOTS (Cross et al., 2019) tracking system, which is an approximate measure of adherence, using data of $\sim 17,000$ patients from Mumbai.

Our work, on the other hand, directly uses the NTEP program data at scale, culled from across locations in a country that has the highest number of TB-afflicted individuals in the world. The data is then used to predict true non-adherence, albeit of an extreme kind, rather than an approximate proxy.

3. Data

Nikshay data. In this work, we use TB patient data obtained via the Nikshay[1] system. Nikshay is a national patient management system for TB control developed and maintained by the Central TB Division (CTD), Government of India, the National Informatics Centre (NIC), and the World Health Organization (WHO). It contains a line list database of TB patients, longitudinally tracking each patient's health status and interactions with the public health system, from diagnosis to treatment completion .

We employ Nikshay data from 2020 of four Indian states: Karnataka, Maharashtra, Uttar Pradesh, and West Bengal. The data (Table 1), being highly sensitive and personal, is anonymized and stripped off of personally identifiable information (patient's name, number, address, etc.). The data consists of separate tables called *registers*, linked by a common patient ID. The 2020 data has seven such registers: Notification, Comorbidity, Contact Tracing, Patient Data, Ad-

1. https://www.nikshay.in/Home/AboutUs

herence, Patient Lab, and DMC (Designated Microscopy Centre) (Table 5). Details of the data analysis and pre-processing pipeline are given in Appendix A.

3.1. Forward splitting

The deployment scenario involves training the model on past data and carrying out inference on future data. In order to simulate this, we forward-split the data in time. We reserve the last six months of data as a *passive evaluation split*, denoted as $(\mathbf{y}^{(\text{pes})}, \mathbf{X}^{(\text{pes})})$, which is only used for final model evaluation and reporting, not for training, optimization, model selection, or internal evaluation. All previous data is referred to as the *modeling split*, $(\mathbf{y}^{(\text{ms})}, \mathbf{X}^{(\text{ms})})$. The modeling split is further divided chronologically in 60 : 20 : 20 proportion into train, validation, and test sets (denoted as $(\mathbf{y}^{(\text{ms,str})}, \mathbf{X}^{(\text{ms, str})})$ with str $\in$ {train, val, test}). The validation set is used, as usual, for hyperparameter optimization within a model class, and the test set is used for selecting the best model class. This best model class is then trained on the entire modeling split prior to its final evaluation on the passive evaluation split.

The 60 : 20 : 20 proportion was decided based on an empirical investigation of different splits to achieve a balance between having sufficient data for training (Section 4) and cohort-wise evaluation (Section 5.4).

Table 1 also summarizes numbers of patients by split. Note a significant drop in the fraction of LFU patients between the modeling and the passive evaluation splits. This can be attributed to the high burden of COVID-19 on the healthcare system during this time period. We expect our models to face such distribution shifts in deployment as well. See Appendix E for effects of distribution shifts on model performance.

4. Modeling

4.1. Encodings

Except for patient age, all features in $\mathbf{X}$ are categorical, and some with high cardinality. Since most models do not work out-of-the-box on such data, we need to encode the features into a real-valued vector space.

We experiment with a number of encoding techniques: those that do not use label information, such as Normalized Count Encoding, Similarity Encoding (Cerda et al., 2018), and Entity Embedding (Guo and Berkhahn, 2016), as well as those that use the label as a prior, such as Target Encoding (Micci-Barreca, 2001), Leave-One-Out (LOO) Encoding, and CatBoost Encoding (Prokhorenkova et al., 2018). We also experiment with recent techniques such as Gap Encoding (Cerda and Varoquaux, 2020a) and MinHash Encoding (Cerda and Varoquaux, 2020a) (to handle high-cardinality).

We compare these encoding techniques (see Figure 2(a) and Table 12) using high-depth XGBoost trees (Chen and Guestrin, 2016), which are convenient for this comparison because they are fast to train and known to do well out of the box on tabular data.

4.2. Metrics

Keeping deployment in mind, the following considerations were found to be relevant to NTEP staff. First, the ML solution should capture as many patients as possible who will eventually be LFU (high recall). Second, the staff workers who will use our solution can typically target only a certain $k\%$ of the patients for intensive monitoring and interventions due to resource limitations and their involvement in other disease programs. Further, there is low prevalence ($\sim 3-4\%$) of the positive class among TB patients. Thus, most standard classification evaluation metrics are ill-suited to the problem. Instead, we focus on the top $k\%$ patients, aiming to

Table 1: State-wise data distribution, split into data for modeling and data for passive evaluation, with number of patients and LFU prevalence.

States	Overall		Modeling Split		Passive Evaluation Split	
	No. of Patients	LFU Rate	No. of Patients	LFU Rate	No. of Patients	LFU Rate
Karnataka	65,120	2.67%	34,171	2.70%	28,041	2.91%
Uttar Pradesh	376,028	3.39%	173,405	4.04%	183,841	3.11%
West Bengal	79,807	2.05%	39,651	2.17%	37,378	2.07%
Maharashtra	157,997	2.51%	75,103	2.98%	75,930	2.28%
Total	678,952	2.96%	322,330	3.42%	325,190	2.78%

achieve high recall in that set, and work with the following two custom metrics:

• Recall@k: Recall for top k% patients when patients are sorted in descending order of the score given by the model. Such metrics are commonly used in ranking problems (Manning, 2008). Feedback from public health experts suggests that TB health workers can typically target roughly 20% of the patients served by their TB unit for interventions. This led us to report recall at $k = 20$ as one of the key metrics.

• AvRecall$(10, 40)$: The average Recall@k for $k \in [10, 40]$, i.e., when targeting the top high risk 10% to 40% of patients, which is a reasonable range for on-ground health worker resource bandwidth, and a more robust measure of model performance on a range of k values.

See Appendix C for further discussion of our metrics.

4.3. Models

Our modeling experiments included simple ML models like decision trees, k-nearest neighbors (k-NN), and naive Bayes classifiers, as well as more recent variants of gradient-boosted decision tree (GBDT) models and deep learning models. We find that all models except k-NN outperform the rule-based baselines (Section 5.1).

GBDT variants include XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018). We also investigated two deep learning architectures that have shown comparable performance to GBDTs on several tabular datasets, TabNet (Arik and Pfister, 2021) and the heavily regularized multilayer perceptron (MLP) (Kadra et al., 2021).

The experiments in Section 5 show that a number of models perform nearly equally well (Table 2). In such situations, (Fisher et al., 2019) suggest that there likely exists a simpler, fully interpretable model with similar performance; we therefore also implement an Explainable Boosting Machine (EBM), a tree-based gradient-boosted generalized additive model (Lou et al., 2013).

Hyperparameter tuning and model selection. We perform hyperparameter tuning, and model and encoder selection using the evaluation metric AvRecall$(10, 40)$ as the loss function $\mathcal{L}$ instead of the training loss.

Our entire hyperparameter tuning and encoder/model selection process can be described as follows:

$$f^* = \arg\min_{f \in \mathcal{F}} \min_{\lambda \in \Lambda_f} \mathcal{L}(\mathbf{y}, f(\mathbf{X}; \lambda, E^*)), \text{ s.t.}$$

$$E^* = \arg\min_{E \in \mathcal{E}} \min_{\lambda \in \Lambda_{f_{\text{xgb}}}} \mathcal{L}(\mathbf{y}, f_{\text{xgb}}(\mathbf{X}; \lambda, E)),$$

where $\mathcal{F}$, Λ_f, and $\mathcal{E}$ denote the set of all models, hyperparameter configurations of model f, and encodings we considered, respectively. We find that the performance of models depends strongly on the encoding of categorical variables, as in (Cerda and Varoquaux, 2020b). With a slight abuse of notation, therefore, we change $f(\mathbf{x})$ to $f(\mathbf{x}; \lambda, E)$,

where λ and E denote model hyperparameters and encoding respectively. While a joint optimization over $\mathcal{F}$ and $\mathcal{E}$ would be ideal, this is practically infeasible due to dataset size and the combinatorial explosion of possibilities. Hence, we initially perform a best encoder search with the XGBoost (f_{xgb}) model, a well-performing GBDT model on tabular data. There is also a manual search over hyperparameters of encoders which is not indicated in the expression.

Optimal model hyperparameters are found via a search over $\lambda \in \Lambda_f$, using a Tree-structured Parzen Estimator approach (TPE) (Bergstra and Bengio, 2012), as implemented in the Hyperopt library (Bergstra et al., 2013). Appendix D lists model details, including hyperparameters and their ranges. We used the validation set $\mathbf{X}^{(\text{ms,val})}$ of the modeling split for hyperparameter optimization, and the corresponding test set $\mathbf{X}^{(\text{ms,test})}$ for the encoder and model search. Once the best model f^* and its best hyperparameter set λ_f^* are found, the model is retrained on the entire $\mathbf{X}^{(\text{ms})}$ set to yield the model $f^{(\text{ms})^*}$. [Note that the best model search also includes an ensemble of five other best models; see Appendix D.2 for details.] For final prediction, we apply $f^{(\text{ms})^*}$ to the passive evaluation set as

$$\mathbf{y}_i^{(\text{pes})} = f^{(\text{ms})^*}(\mathbf{x}_i^{(\text{pes})}; \lambda_f^*, E^*),$$

or on its various cohorts as detailed in Section 5.4.

5. Results and Evaluation

5.1. Rule-based baselines

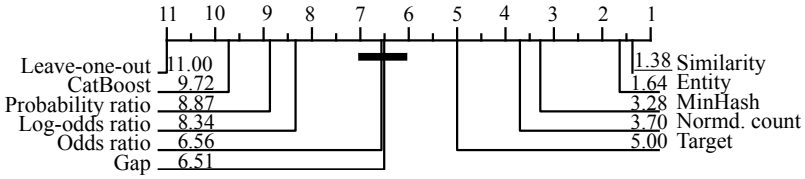
We consider three rule-based heuristics as benchmarks to compare the performance of our ML models. These rules are based on correlations in the data, findings by the CTD, protocols suggested by the Karnataka Health Promotion Trust (KHPT) (KHPT, 2019), the TB ground staff, and previous literature vetted by domain experts (Cherkaoui et al., 2014; Jaggarajamma et al., 2007; Bhatnagar, 2019; Bhattacharya et al., 2018). These baselines are occasionally used by health workers to prioritize interventions. We list the baselines with their associated features and performance in Appendix B. Overall, we find that even the best rule-based baseline suffers from poor LFU recall.

5.2. Statistical comparisons across models and encodings

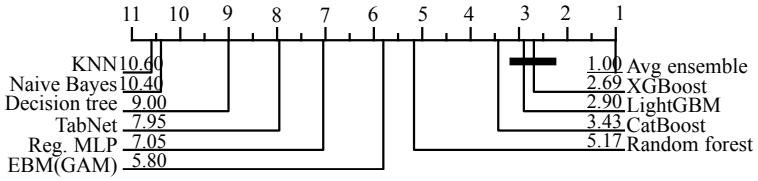
We use a Critical Difference (CD) diagram (Demšar, 2006) to represent the best encoder E^* and model f^* with the modeling split of the data (Figure 2). The CD diagram is typically used to compare multiple models over multiple data sets statistically. To plot this, we collect 1000 bootstrap samples from the test set $\mathbf{X}^{(\text{ms,test})}$, each of the same size as the test set. Then we run different encoders or models on each of these bootstrap samples yielding 1000 values for AvRecall(10,40). The Friedman test is first performed to reject the null hypothesis (all models are statistically equivalent), followed by a post-hoc analysis based on the Wilcoxon significance test, in which critical differences are calculated. Further details on bootstrapping are given in Appendix I.

5.3. Results on modeling split

Most ($\sim 98\%$) features are categorical in nature. Figure 2 and Table 12 illustrate that for voluminous tabular data with mainly categorical features, some of which are of high cardinality, performance is quite sensitive to the encoding used. We find that similarity encoding performs the best among the extensive list of encoders we considered. See Appendix F a for detailed comparision. Figure 2(*b*) shows a comparison of results of all models in terms of their AvRecall(10,40) performance using similarity encoding. Table 2 lists appropriate metrics for these models; we find that the XGBoost model outperforms other individual models, with Light-



(*a*) Selecting best encoder



(*b*) Selecting best model

Figure 2: **Critical difference diagrams** with a Wilcoxon significance analysis on 1000 bootstrap samples. The numbers on the top line indicate average ranks, and connected ranks via a bold bar indicate that performances are not significantly different ($p > 0.05$).

GBM and CatBoost being very close. Despite the recently reported encouraging performance of deep learning models on tabular data (Arik and Pfister, 2021; Kadra et al., 2021), we note that GBDT models still robustly outperform them on our large real-world tabular dataset (as noted in (Shwartz-Ziv and Armon, 2022) as well).

Note that, while Table 2 displays the lift in ML model performance over the best baseline, the lift over random selection of patients for targeting is much more spectacular. For a random classifier, Recall@k is $k/100$ and AvRecall(10,40) is 0.25. Hence, for the best model in Table 2 (the average ensemble model), the percentage lift is $\sim$ 214% (Recall@20) and $\sim$ 172% (AvRecall(10,40)) from a random classifier. This is especially relevant because TB field staff currently are not known to follow specific standardized rules for patient prioritization.

We also find, as shown in Table 2, that the average ensemble of our top-5 best models slightly improves metrics over XGBoost. This may be attributed to relatively low output correlation between EBM and the other models (Appendix D.2). On the other hand, we find that the solution is also quite robust in the sense of low predictive multiplicity across models (Appendix C.2). See Appendix D for further modeling details.

Table 2: Performance of various ML models on the test set of the modeling split with similarity encoding with 95% confidence intervals generated from bootstrap samples. "Lift" is percentage increase in Recall@20 over best baseline.

Model	AvRecall(10,40)	Recall@20	Lift(%)
Random Forest	0.658 ± 0.019	0.613 ± 0.023	95.22
k-NN	0.481 ± 0.021	0.425 ± 0.024	35.35
Decision Tree	0.515 ± 0.018	0.469 ± 0.023	49.36
Naive Bayes	0.481 ± 0.018	0.408 ± 0.022	29.94
XGBoost	**0.671 ± 0.019**	**0.624 ± 0.023**	**98.73**
CatBoost	0.668 ± 0.020	0.625 ± 0.024	99.04
LightGBM	0.670 ± 0.020	0.620 ± 0.023	97.45
Regularized MLP	0.612 ± 0.021	0.566 ± 0.024	80.25
TabNet	0.602 ± 0.02	0.564 ± 0.023	79.62
EBM (GAM)	0.650 ± 0.020	0.606 ± 0.025	92.99
Avg. Ensemble	**0.681 ± 0.019**	**0.627 ± 0.025**	**99.68**

5.4. Results on passive evaluation split

We apply the best individual and ensemble trained models selected in the previous step to the passive evaluation holdout set $(\mathbf{y}^{(\text{pes})}, \mathbf{X}^{(\text{pes})})$. We find that the XGBoost model applied to this set yields a Recall@20 of **0.614** and AvRecall(10,40) of **0.658**, while the average ensemble model yields Recall@20 of **0.618** and AvRecall(10,40) of **0.663**. At 20% patient targeting, assum-

ing that all patients identified as at risk by the model are subjected to intensive interventions by health workers, this corresponds to saving 5587 patients from becoming LFU over a six-month period across four Indian states, instead of 1808 and 2974 patients via the random classifier and best rule-based baseline yielding a lift of $\sim 209\%$ and $\sim 88\%$ respectively.

Table 3 displays the passive evaluation results broken down by month, signifying the generalization capabilities of the model trained on earlier data. There is no significant reduction in performance by month. Generalization capabilities across datasets spanning longer time periods are detailed in Appendix E. Furthermore, Table 3 shows that the drop in performance on the passive evaluation time split relative to the modeling split test set (which was used to optimize model selection) (Table 2) is not appreciable.

We also carried out a detailed evaluation study of our best ensemble model across different data cohorts. For this, we use a global score threshold identified from the top-20% of patients in the entire passive evaluation dataset, and an alternative to this global threshold by implementing a local district-level threshold (Appendix G).

Table 3: Passive evaluation results for "Month of TB Treatment Initiation" for 2020.

Month	#Patients	LFU Rate	Recall@20
Jul	48,208	3.18%	0.634 ± 0.024
Aug	42,797	3.02%	0.604 ± 0.025
Sep	49,759	3.05%	0.629 ± 0.024
Oct	51,639	2.82%	0.617 ± 0.025
Nov	50,288	2.80%	0.614 ± 0.014
Dec	62,128	2.51%	0.605 ± 0.025

We consider two types of cohorts in our detailed evaluation.

Geographical cohorts: We analyze performance at the state and district levels. Beyond measuring robustness of the model to geographical variations, this information is particularly useful in selecting geographies for deploying a pilot based on the model.

Non-geographical cohorts: These cohorts are selected based on personal features, such as gender; temporal variables such as the month of treatment initiation and the time elapsed till the final outcome; and TB program-relevant cohorts such as type of healthcare system (public or private care) and type of the disease (drug-sensitive vs drug-resistant TB). We observe that our model performs exceptionally well on drug-resistant patients (the sub-group with the highest mortality rate) with a Recall@20 of **0.924** (Table 15).

The full results of cohort-based evaluation in Appendix G show that model performance is robust with respect to most cohorts of interest, except geographical cohorts and, to a lesser extent, gender. We discuss means of mitigating low performance in underperforming cohorts in Section 5.6 below.

5.5. Interpretability

Interpretability is an important consideration in ML solutions for healthcare, both in terms of accountability and stakeholder involvement. For our use case, understanding important features is critical to designing effective interventions. To this end, we implement Permutation Feature Importance (PFI) (Breiman, 2001), a global interpretability method, with respect to the Recall@20 metric (Table 4, with details in Appendix H). The importance of the covariate *BankDetailsAdded* for model performance is interesting. This feature is in fact a proxy for patients on recurring monthly government monetary benefit schemes throughout their treatment course. This continuous positive reinforcement decreases LFU probability.

We also implement the Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) method with a linear regressor for patient-level explanations, which also identifies BankDetailsAdded as a majority top feature. Appendix H has further details and examples.

Table 4: Top-10 feature importance values with respect to Recall@20 using Permutation Feature Importance (PFI). More details in Appendix H.1.

Feature	Importance	Std. Dev.
BankDetailsAdded	0.105	0.002
ReasonforTesting	0.060	0.002
CurrentFacilityPHIType	0.029	0.001
CurrentFacilityTBU	0.026	0.001
CurrentFacilityDistrict	0.017	0.001
CurrentFacilityPHI	0.015	0.002
resultSampleA	0.012	0.001
Age	0.009	0.002
TypeOfCase	0.008	0.001
ContactTracing_Done	0.008	0.001

5.6. Improving performance on underperforming cohorts

The model does reasonably well across most cohorts (see Section G), but has variation across geographies and gender. We explore two approaches towards improving performance on underperforming cohorts and reducing variance; these are critical for our planned country-wide solution deployment.

- *Data augmentation*: Augmenting the training data with copies of data from an underperfoming cohort, similar to increasing the effective weight of patients from that cohort in the loss function, cohort. We do this at the district level, as discussed in Appendix J.1.
- *Algorithmic fairness*: Algorithmic fairness is a critical consideration in AI for healthcare settings. We investigate a post-hoc fairness approach (Rodolfa et al., 2021) that adjusts model scores to equalize our metric across important cohorts such as states, districts and gender, as detailed in Appendix J.2.

Both approaches significantly improve underperforming cohorts, especially the very low-scoring ones, with little impact on better ones, and a reduction in the overall variance. For instance, districts with Recall@20 less than 0.3 show an 87% lift in mean Recall@20 with augmentation, and 132% with algorithmic fairness, albeit with an increase in effective k for the latter (computed from Table 19 and Table 21). For gender, the fairness method also gives a modest improvement on the "Female" cohort (Table 22).

6. Conclusions

This work describes the development of an ML solution for the prediction of LFU in TB patients. India, with the world's largest TB burden, and possessing one of the world's largest centrally administered longitudinal databases for tracking TB patients, represents a natural ground for testing and deployment of such solutions. The models were developed keeping in mind the complex deployment scenario, with data issues, distributional drift, and high cardinality features. We showed that a solution based on an ensemble combining the five best models is the best-performing across model classes (verified statistically) and important cohorts, and generalizes well across time. We explicitly address interpretability concerns and performance improvement using data augmentation and algorithmic fairness, and find that the solution works especially well on DRTB patients, one of the most vulnerable patient cohorts.

Future technical directions include end-to-end learnable embeddings and fully interpretable modeling such as Neural Basis Models (NBM) (Radenovic et al., 2022). We are looking at extending to other adverse outcomes such as pre-treatment loss to follow up, treatment failure and mortality, potentially in a multiclass setting. Beyond classificiation, prediction of time-to-LFU is a richer, more useful, but harder problem to solve.

We hope that the results and insights from this work can be helpful to others working on similar problems, and that its deployment can mark a successful milestone in the long journey towards tuberculosis eradication.

References

J Adams and Jan Scott. Predicting medication adherence in severe mental disorders. *Acta Psychiatrica Scandinavica*, 101, 2000.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4): 1059–1086, 2020.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *AAAI*, volume 35, pages 6679–6687, 2021.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13 (2), 2012.

James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*, volume 13, page 20. Citeseer, 2013.

Himani Bhatnagar. User-experience and patient satisfaction with quality of tuberculosis care in india: a mixed-methods literature review. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, 17:100127, 2019.

Tridibes Bhattacharya, Soumalya Ray, Puspendu Biswas, and D Das. Barriers to treatment adherence of tuberculosis patients: A qualitative study in west bengal, india. *International Journal of Medical Science and Public Health*, 7(5):1, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Patricio Cerda and Gaël Varoquaux. Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, 2020a.

Patricio Cerda and Gaël Varoquaux. Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, 2020b.

Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8):1477–1494, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *SIGKDD*, pages 785–794, 2016.

Imad Cherkaoui, Radia Sabouni, Darya Kizub, Alexander C Billioux, Kenza Bennani, Jamal Eddine Bourkadi, Abderrahmane Benmamoun, Ouafae Lahlou, Rajae El Aouad, and Kelly E Dooley. Treatment default amongst patients with tuberculosis in urban morocco: predicting and explaining default and post-default sputum smear and drug susceptibility results. *PloS one*, 9(4):e93574, 2014.

Andrew Cross, Nakull Gupta, Brandon Liu, Vineet Nair, Abhishek Kumar, Reena Kuttan, Priyanka Ivatury, Amy Chen, Kshama Lakshman, Rashmi Rodrigues, et al. 99dots: a low-cost approach to monitoring and improving medication adherence. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, pages 1–12, 2019.

Estella M Davis, Kathleen A Packard, and Cynthia A Jackevicius. The pharmacist role in predicting and improving medication adherence in heart failure patients. *Journal of Managed Care Pharmacy*, 2014.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Mark L Ettenhofer, Jessica Foley, Steven A Castellon, and Charles H Hinkin. Reciprocal prediction of medication adherence and neurocognition in hiv/aids. *Neurology*, 2010.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

K Jaggarajamma, G Sudha, Vendachalam Chandrasekaran, Charles Nirupa, A Thomas, Thottikkamath Santha, Malaisamy Muniyandi, and PR Narayanan. Reasons for non-compliance among patients treated under revised national tuberculosis control programme (rntcp), tiruvallur district, south india. *Indian Journal of Tuberculosis*, 54 (3):130–135, 2007.

Amita Jain and Pratima Dixit. Multidrug resistant to extensively drug resistant tuberculosis: what is next? *Journal of Biosciences*, 33(4):605–616, 2008.

Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. 2021.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. volume 30, 2017.

India KHPT. Differentiated care model, standard operating procedures: A personalized approach to prevention, care and support for tb patients, 2019. http://www.khpt.org/wp-content/uploads/2019/03/DCM-SOP-with-RANA.pdf.

Jackson A Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra Dilkina, and Milind Tambe. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *SIGKDD*, 2019.

Kai Kliiman and A Altraja. Predictors and mortality associated with treatment default in pulmonary tuberculosis. *The international journal of tuberculosis and lung disease*, 14, 2010.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *SIGKDD*, pages 623–631, 2013.

Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *ICML*, 2020.

Habtamu Sewunet Mekonnen and Abere Woretaw Azagew. Non-adherence to anti-tuberculosis treatment, reasons and associated factors among tb patients attending at gondar town health centers, northwest ethiopia. *BMC Research Notes*, 11(1):1–8, 2018.

Daniele Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1):27–32, 2001.

Government of India MoHFW. Tuberculosis report. *India TB Report*, 2022. https://tbcindia.gov.in/index1.php?lang=1&level=1&sublinkid=5613&lid=3658.

Salla A Munro, Simon A Lewin, Helen J Smith, Mark E Engel, Atle Fretheim, and Jimmy Volmink. Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS Medicine*, 4 (7):e238, 2007.

Muhammad Osman, Sue-Ann Meehan, Arne von Delft, Karen Du Preez, Rory Dunbar, Florian M Marx, Andrew Boulle, Alex Welte, Pren Naidoo, and Anneke C Hesseling. Early mortality in tuberculosis patients initially lost to follow up following diagnosis in provincial hospitals and primary health care facilities in western cape, south africa. *PLoS One*, 16(6):e0252084, 2021.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *NIPS*, 31, 2018.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. *arXiv preprint arXiv:2205.14120*, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144, 2016.

Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3, 2021.

Nirmalya Roy, Mausumi Basu, Sibasis Das, Amitava Mandal, Debashis Dutt, and Samir Dasgupta. Risk factors associated with default among tuberculosis patients in darjeeling district of west bengal, india. *Journal of Family Medicine and Primary Care*, 4, 2015.

Estifanos Biru Shargie and Bernt Lindtjørn. Determinants of treatment adherence among smear-positive pulmonary tuberculosis patients in southern ethiopia. *PLoS medicine*, 4(2):e37, 2007.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Jimmy Volmink and Paul Garner. Directly observed therapy for treating tuberculosis. *Cochrane Database of systematic reviews*, 4, 2007.

Reynold Washington, Rajaram Subramanian Potty, A Rajesham, T Seenappa, Anil Singarajipura, Reuben Swamickan, Amar Shah, KH Prakash, Arin Kar, Karthikeyan Kumaraswamy, et al. Is a differentiated care model needed for patients with tb? a cohort analysis of risk factors contributing to unfavourable outcomes among tb patients in two states in south india. *BMC public health*, 20(1):1–12, 2020.

Geneva: World Health Organization WHO. Global tuberculosis report 2021, 2021.

Minlan Xu, Urban Markström, Juncheng Lyu, and Lingzhong Xu. Detection of low adherence in rural tuberculosis patients in china: application of morisky medication adherence scale. *International Journal of Environmental Research and Public Health*, 2017.

Zhigang Zheng, Eric J Nehl, Chongxing Zhou, Jianjun Li, Zhouhua Xie, Zijun Zhou, and Hao Liang. Insufficient tuberculosis treatment leads to earlier and higher mortality in individuals co-infected with hiv in southern china: a cohort study. *BMC infectious diseases*, 20(1):1–10, 2020.

Appendix A. Data Quality and Preprocessing

A brief description of the registers available in the Nikshay database is given in Table 5, and the percentage of missing values in the data for each state is given in Table 6.

A number of features in the data are categorical, with some of them having high cardinality. Thus, they are not fit for direct modeling, which motivates us to experiment with a large number of encoding techniques. Also, there are a number of data quality issues such as missing values, misspelled fields, etc., in several columns. We clean and transform the data in multiple steps. First, we remove features with a high proportion of missing or unintelligible string values. We then use basic imputation techniques (mean imputation, creating an additional "missing" category) to fill in missing and invalid values in the remaining features. Finally, we merge all the registers based on a hashed EpisodeID key (which acts as the primary key since a TB patient can experience multiple episodes of TB). We also consult domain experts to ensure that the categories have consistently named strings across registers and that there is no retrospective data in the input that can cause leakage.

Appendix B. Rule-based Baselines

Table 7 lists features included in the three rule-based baseline models. The baseline scores are defined to be the fraction of features that lie within a specified range. Rule 1 comes from NTEP draft guidelines (unpublished), Rule 2 from guidelines proposed by (KHPT, 2019), while the best-performing baseline (Rule 3) comes from an extensive survey of the literature (Jaggarajamma et al., 2007; Bhatnagar, 2019; Bhattacharya et al., 2018), and analysis of features that correlate the most with the LFU outcome in the Nikshay data.

Appendix C. Metric Details

C.1. Recall@k

We note that Recall@k has an upper bound of $\min(1, k/p)$, where p is the prevalence of the positive class. This bound, however, is not of concern for us since the prevalence is considerably less than the percentage targeted. A model that assigns random risk scores from $[0, 1]$ to all patients is expected to have a Recall@20 of 0.20.

While the value $k = 20$ for Recall@k and the range $k \in (10, 40)$ for AvRecall are anecdotal, Figure 3 shows that our models, although trained to optimize over AvRecall(10,40), perform well for all values of k.

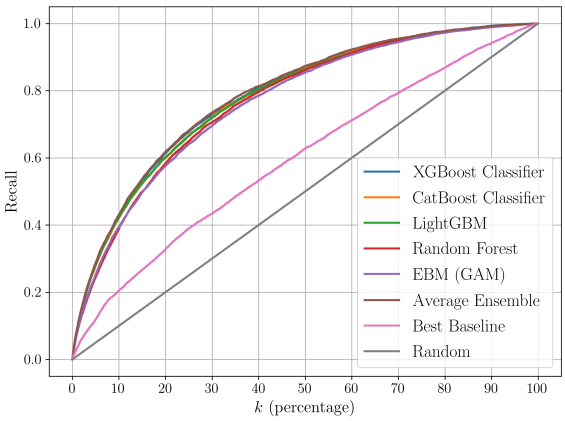


Figure 3: Recall@k vs k for our best models, best baseline and random classifier on the passive evaluation split.

C.2. Ambiguity and discrepancy

We see in Table 2 that a number of models perform similarly on our aggregated, rank-based metric. However, on individual patients, it is possible that these models give very different scores. Predictive multiplicity is defined as the ability of a prediction problem to admit competing models with conflicting predictions. For evaluating the robustness of our solution against model selection and improving stakeholder participation, we

Table 5: Registers used to store patient-level data in Nikshay and the number of features from the registers used by the models. Features with low prevalence and potential leakage were removed during the pre-processing phase.

Register	Meaning	#Features
Notification	Basic details about the patient collected at notification time such as age, gender, location, health facility type, TB type, etc.	54
Comorbidity	Comorbidity information such as HIV and diabetes status, tobacco and alcohol addiction, etc.	17
Patient Data	Demographic details such as occupation, marital status, etc.	3
Adherence	Information about adherence technology used	1
Patient Lab	Lab results such as CBNAAT, microscopy, culture, chest X-ray.	1
DMC	Results from Designated Microscopy Centre (DMC)	3

Table 6: State-wise percentage of missing/null values for all covariates across all patients in 2020.

State	#Patients	Null Values
Karnataka	65,120	6.83%
Uttar Pradesh	3,76,028	7.62%
West Bengal	79,807	10.71%
Maharashtra	1,57,997	6.85%
Total	6,78,952	9.17%

compute two metrics of predictive multiplicity (Marx et al., 2020) for the set of our best-performing models. Let h_0 be the best performing model, with performance metric p, and $S_\epsilon(h_0)$ be a set of models with a performance range of $[(1-\epsilon)p, p]$. we define Ambiguity (α_ϵ) and Discrepancy (δ_ϵ) as follows:

$$\alpha_\epsilon(S_\epsilon) := \frac{1}{n}\sum_{i=1}^{n} \max_{h \in S_\epsilon(h_0)} \mathbb{1}[h(\mathbf{x_i}) \neq h_0(\mathbf{x_i})],$$

$$\delta_\epsilon(S_\epsilon) := \frac{1}{n} \max_{h \in S_\epsilon(h_0)} \sum_{i=1}^{n} \mathbb{1}[h(\mathbf{x_i}) \neq h_0(\mathbf{x_i})],$$

where n is the number of patients. We take an $\epsilon = 0.2$ set of similarly performing models, and find that they have an ambiguity score of 22.24% and a discrepancy score of 13.51% on the test set of the modeling split. Comparing with the same measures on other real-world datasets from (Marx et al., 2020), we find these values well within the acceptable range.

Appendix D. Modeling Details

D.1. Hyperparameter search spaces

We run a 100-iteration Hyperopt hyperparameter search on every model. The hyperparameter search space used for XGBoost models is given in Table 8.

For the regularized MLP model, as suggested by (Kadra et al., 2021), we create an MLP with 5 fully-connected layers with a skip connection from the second layer to the last layer. Hyperparameters such as the size of the layers, the learning rate, the optimizer, the weight of the low-prevalence class in the loss function, dropout, SGD momentum, gradient clip, early stopping, learning-rate scheduling, etc. are optimized through TPE. Further details for all models are made available through a shared repository.[2]

D.2. Ensembling

On analyzing model scores, we notice a large correlation between the GBDTs, which all have a relatively low correlation with EBM as shown in Figure 4. To try and exploit this, we investigate ensembling EBM with XGBoost, Catboost, LightGBM, and Ran-

2. For the complete list of hyperparameter search spaces for all our models, please visit: https://anonymous.4open.science/r/model_configurations-B003/

Table 7: Rule-based baseline models, and their performance on the test set of the modeling split.

Baseline	**Features**	**Recall@20**
Rule 1	Alcohol intake history (present), HIV status (positive), Diabetes status (positive)	0.184
Rule 2	Type of Case (DRTB or retreatment), HIV status (positive), Diabetes Status (positive), Age ($\geq$ 60), Household contacts above age six (absent)	0.282
Rule 3	Age ($\in$ [18, 45]), Type of Case (DRTB or retreatment), HIV Status (positive), Diabetes Status (positive), Gender (male or transgender), Migrant:change of district after diagnosis (yes) , Alcohol intake history (present), Household contacts above age six (absent), Microbiologically Confirmed (yes), Disease site (pulmonary), UDST Done (yes), Bank Details Added (no)	0.314

Table 8: The hyperparameter search space for XGBoost models. Optimization was done using TPE on the AvRecall(10, 40) values on the validation set of the modeling split.

Hyperparameter	**Distribution**	**Range**
Learning rate	Log Uniform	$[10^{-7}, 1]$
Num. estimators	Discrete Uniform	$\{50, 100, 200, 400, 600, 800, 1000, 1200, 1500, 2000\}$
Maximum depth	Uniform Integer	$[1, 9]$
Min child weight	Uniform Integer	$[1, 8]$
Scale pos. weight	Uniform Integer	$[1, 90]$
L1 term	Log Uniform	$[10^{-5}, 1]$
L2 term	Uniform	$[10^{-3}, 1]$
Subsample ratio	Uniform	$[0.5, 1]$
Col sample ratio	Uniform	$[0.5, 1]$
Min split loss	Discrete Uniform	$\{0, 0.3, 0.6, 0.9, 1.2, 2, 4, 6, 8\}$

dom Forest. An averaged ensemble leads to a small boost in performance (Table 2).

Appendix E. Temporal Generalization Analysis

We carried out various experiments to test the ability of the model to generalize across time.

First, we initially obtained a smaller dataset of patients from the year 2019 in order to establish a proof of concept. We split this into a training set of 38,000 patients and a holdout set of 3,800 patients (this evaluation is termed PoC-2019). We later gained access to all of the 2019 data and 2020 data. We received access to four registers – notification, comorbidity, contact tracing, and patient lab registers – for the year 2019, while for 2020 data, three additional registers were shared. As described in Section 3.1, the data was split temporally in a forward fashion to mimic actual deployment. We performed passive evaluation on these datasets (termed PE-2019 and PE-2020). Results from these evaluations are summarized in Table 9, and show good temporal generalization.

We then designed the following experiments with our best model class from 2019, using the four registers available from then:

Experiment 1: Best model trained with the first six months of the 2019 data and evaluated separately on the last six months of 2019 and the first six months of 2020.

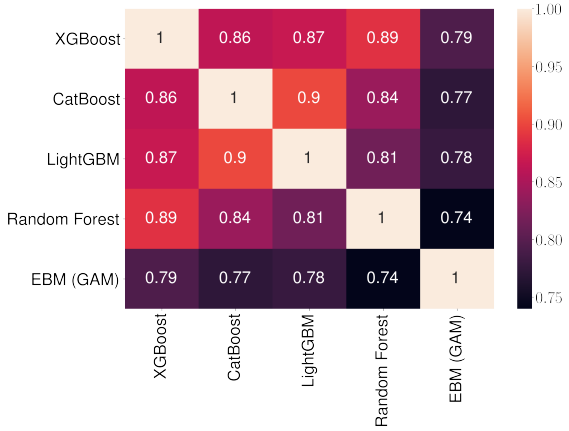


Figure 4: Correlation matrix for our top-5 models with similarity encoding on the passive evaluation split.

Table 9: Summary of results on different datasets received. We note that Recall@20 is roughly constant across these evaluations. The superscript to n indicates the data split.

Evaluation	$n^{(ms)}$	$n^{(pes)}$	Recall@20
PoC-2019	38,243	3800	0.581
PE-2019	943,654	431,932	0.569
PE-2020	678,952	325,190	0.618

Experiment 2.1: Model trained on all of the 2019 data and tested on the sixth month of the 2020 data.

Experiment 2.2: Model trained on 2019 and the first five months of 2020 data and tested on the sixth month of 2020 data.

The aim of Experiment 1 is to understand how robust a previously trained model is on new data, and that of Experiments 2.1 and 2.2 is to understand the importance of the availability of immediately preceding data in the performance of the model. The result of Experiment 1 shows that the model is inherently robust and generalizes well across time as the Recall@20 value remained consistent from 2019 to 2020 (Table 10). On the other hand, the results of Experiments 2.1 and 2.2 suggest a significant increase in performance when the model is trained using the immediately preceding data (Table 11).

Table 10: Comparison of Recall@20 in Experiment 1, showing the generalization ability of the model. Training on the first half of 2019 data leads to similar performance on the second half of 2019 data and the first half of 2020 data.

Experiment 1	2019	2020
Recall@20	0.560	0.578

Table 11: Comparison of Recall@20 between Experiments 2.1 and 2.2. For the same test set, a model that is trained on the immediately preceding data outperforms a model that is trained much earlier.

Experiment 2	Test Set Recall@20
Experiment 2.1	0.572
Experiment 2.2	0.606

We observe that the model is quite robust on later data and also that the addition of the latest data helps it adapt to distribution shifts.

Appendix F. Comparison of Encodings

Here, we flesh out the findings of Section 5.2 with respect to comparison of encodings. Table 12 shows the performance of various encoding schemes we used in the work. The uncertainty is estimated using a 95% confidence interval of the results from bootstrap samples, as described in Section 5.2 and Appendix I.

Table 12: Performance of different encoding and embedding techniques for categorical data using the XGBoost classifier. "Lift" denotes percentage increase in Recall@20 over the best baseline.

Encoding Type	AvRecall(10,40)	Recall@20	Lift
Leave-One-Out	0.373 ± 0.020	0.322 ± 0.022	6.05%
Probability Ratio	0.535 ± 0.020	0.491 ± 0.023	56.37%
Log-Odds Ratio	0.538 ± 0.021	0.502 ± 0.022	59.87%
CatBoost Encoder	0.523 ± 0.022	0.494 ± 0.024	57.32%
Odds Ratio	0.559 ± 0.021	0.524 ± 0.023	66.88%
Target Encoder	0.603 ± 0.020	0.558 ± 0.023	77.71%
Normalized Count	0.653 ± 0.020	0.614 ± 0.024	95.54%
Entity Embedding	0.670 ± 0.020	0.623 ± 0.025	98.41%
Similarity Encoder	**0.671 ± 0.019**	**0.624 ± 0.023**	**98.73%**
Gap Encoder	0.561 ± 0.021	0.508 ± 0.024	61.78%
MinHash Encoder	0.657 ± 0.020	0.614 ± 0.024	95.54%

Appendix G. Cohort-Wise Evaluation of Model Performance

We classify patients into two categories based on the average ensemble model score: high-risk and low-risk, with the corresponding thresholds calculated in accordance with the rank-based metric Recall@20. The top 20 percentile of the patients based on the model scores are classified as high-risk, with the remaining classified as low-risk. We explore two methods of thresholding:

- *Global*: This is a single threshold corresponding to the top-20 percentile of all the patients being evaluated, which is applied universally.
- *Local*: We use a local district-level threshold to ensure that every district has a consistent 20% patients targeted. Thresholds thus calculated are per district and are applied to patients in those districts only.

We define effective k as the percentage of patients targeted in a particular cohort, which would be the k corresponding to using a cohort-level threshold. While a global threshold leads to higher overall sensitivity, it results in different levels of targeting at different districts. This leads to insufficient targeting and poor sensitivity with low effective k in some districts, and an impractically high effective k in some others. We use local thresholding to fix these issues, and consider it more relevant for an on-ground deployment. We also consider using local thresholds at even more granular (sub-district) levels but do not proceed with it because of small population sizes for these cohorts leading to a large variance in the predictions.

While we observe similar performance across both the thresholding methods, we also analyze how they affect the performance in some major cohorts. A few cohorts of importance that we assess are:

- The month of treatment initiation
- Gender
- Type of Case (drug-sensitive vs drug-resistant)
- Peripheral health institution (PHI) type (public vs private)

We observe that the effective k (targeted %) is roughly constant across the six different months of evaluation, as reported in Table 17, indicating robust performance across

time when using a global threshold. However, we also observe that there is a large difference in effective k across the 4 states as shown in Table 13. This leads to a large difference in Recall@20 across these states. On the other hand, local thresholding enables the model to target every state equally and thus boosts the performance in underperforming states (Table 13). We also study the model performance and implications of threshold type used on all non-location cohorts.

For example, performance of the model stays largely constant on the "Gender" cohort (Table 14). We observe that there is an overall improvement in the performance of the model on the "Male" and "Female" cohorts if we use a local threshold. It does, however, lead to a trade-off with a drop in performance on the "Transgender" category. We observe a similar trade-off with the "Type Of Case" cohorts (Table 15). We find that the model has a high sensitivity for LFU prediction on patients who are diagnosed with DRTB (Drug-resistant TB) as compared to DSTB (Drug-sensitive TB patients) in both cases. Using a local threshold results in an increase in performance across the DSTB patients, but that is traded-off by a decrease in performance across the DRTB patients. Since DRTB patients have a higher propensity to become LFU, and have a separate, longer treatment regimen with more intensive care, this is particularly relevant when choosing between the two thresholding methods. On the other hand, we don't observe such trade-offs in the "PHI Type" cohort. This cohort is a useful axis of measurement since the public and private systems are both important while having different ways of working. We observe that local thresholding outperforms global thresholding (Table 16) across both its categories.

While our model performs satisfactorily on all non-location cohorts, we observe that there are multiple trade-offs that need to be considered while choosing between the two thresholding methods. On-ground personnel can, at the district-level, implement further granular thresholding across non-location cohorts (Eg., "Gender") to either improve or equalize the performance across that cohort.

Appendix H. Interpretability Methods

H.1. Permutation Feature Importance (PFI)

Table 18 shows the top-10 features determined by the PFI method on the best model, along with their definitions.

H.2. Accumulated Local Effects (ALE)

We implement global interpretability techniques such as an Accumulated Local Effects (ALE) (Apley and Zhu, 2020) plot to visually determine the effect of each feature on the predicted target. The ALE Plot for the feature "Age" can be viewed in Figure 5. We observe that there is a steep increase in risk as the age increases from 18 to 30, followed by a similar performance till the age of 60 and then a gradual increase in risk beyond age 60. This is consistent with reports from domain experts and helpful in creating appropriate brackets for age-specific interventions.

H.3. Local Interpretable Model-agnostic Explanation (LIME)

The LIME result for a randomly selected patient can be viewed in Figure 6. The patient has a model score of 0.037 and a ground truth label of 0. We gain valuable insights into the model behavior by analyzing the LIME output. Features with positive weights (green) help the model to push the decision towards LFU, while those with negative weights (red) help push it towards non-

Table 13: Passive evaluation results for the location cohort "State" with a global threshold vs a local district-level threshold.

State	Effective k (Global)	Recall@20 (Global)	Recall@20 (Local)
Karnataka	11.981 ± 0.306	0.516 ± 0.020	0.668 ± 0.027
Maharashtra	14.224 ± 0.192	0.607 ± 0.013	0.682 ± 0.013
Utar Pradesh	25.387 ± 0.149	0.669 ± 0.006	0.614 ± 0.008
West Bengal	10.954 ± 0.285	0.403 ± 0.019	0.567 ± 0.014

Table 14: Passive evaluation results for the "Gender" cohort with a global threshold vs a local district-level threshold.

Gender	#Patients	#LFUs	Recall@20 (Global)	Recall@20 (Local)
Female	130,068	3,201	0.563 ± 0.008	0.576 ± 0.014
Male	181,277	5,837	0.645 ± 0.007	0.658 ± 0.005
Transgender	137	4	0.674 ± 0.326	0.552 ± 0.219

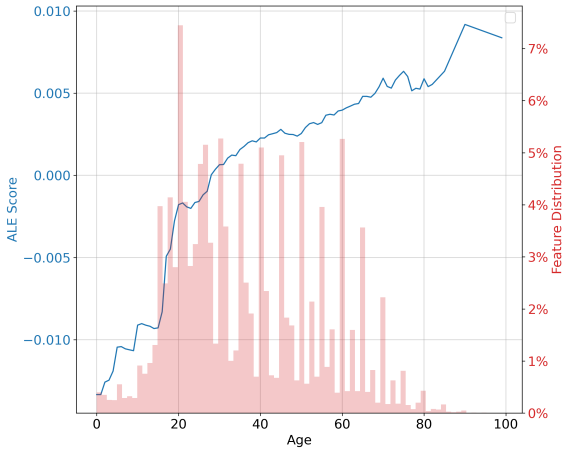


Figure 5: ALE Plot of the feature "Age" with respect to the model outcome.

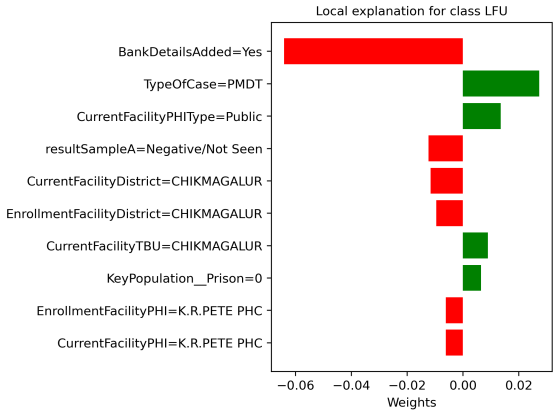


Figure 6: Feature importance values for a randomly selected patient prediction generated using LIME with a Linear Regressor. The feature importances show the impact that various features to give an output of 0.037 for a non-LFU Patient.

LFU. We see that *BankDetailsAdded*, which is a proxy for whether the patient is receiving direct cash transfers, plays a major role in reducing the score of the patient, and that several location features have an impact on the prediction.

Appendix I. Uncertainty Calculation

We use the following algorithm to generate bootstrapped samples:

Table 15: Passive evaluation results for the "Type of Case" cohort with a global threshold vs a local district-level threshold.

Type Of Case	#Patients	#LFUs	Recall@20 (Global)	Recall@20 (Local)
Drug-resistant TB	10,267	909	0.924 ± 0.011	0.907 ± 0.013
Drug-sensitive TB	301,215	8,133	0.586 ± 0.007	0.601 ± 0.008

Table 16: Passive evaluation results for the "PHI Type" cohort with a global threshold vs a local district-level threshold.

PHI Type	#Patients	#LFUs	Recall@20 (Global)	Recall@20 (Local)
Private	87,910	2,443	0.598 ± 0.009	0.610 ± 0.015
Public	223,572	6,599	0.626 ± 0.008	0.641 ± 0.006

Table 17: Passive evaluation results for the cohort "Month of TB Treatment Initiation", for the year 2020, with effective k

Month	Effective k	Recall@20 (Global)
Jul	19.919 ± 0.358	0.634 ± 0.024
Aug	19.579 ± 0.359	0.604 ± 0.025
Sep	20.091 ± 0.364	0.629 ± 0.024
Oct	20.092 ± 0.347	0.617 ± 0.025
Nov	20.976 ± 0.347	0.614 ± 0.014
Dec	19.572 ± 0.319	0.605 ± 0.025

- For a cohort C of size N, draw 1000 N-sized samples, $s_1, \ldots, s_{1000}$ with replacement from C.
- Run the model on the s_i's to generate 1000 N-sized lists of model scores $l_1, \ldots, l_{1000}$.
- Compute our metric of interest (e.g. Recall@20, AvRecall$(10, 40)$) on the l_i's, generating 1000 samples of our metric $m_1, \ldots, m_{1000}$.

The m_i's can be used for generating uncertainty intervals, boxplots (see Figure 8 for models and Figure 9 for encoders), critical diagrams, etc. Note that the training set is fixed throughout this pipeline and we do not use bootstrapping during model selection (on the val set) either. See Figure 7 for an example of generated uncertainty intervals using bootstrap, using confidence intervals obtained from the empirical distribution and a fitted Gaussian. We note that 95% confidence intervals for both methods are nearly identical.

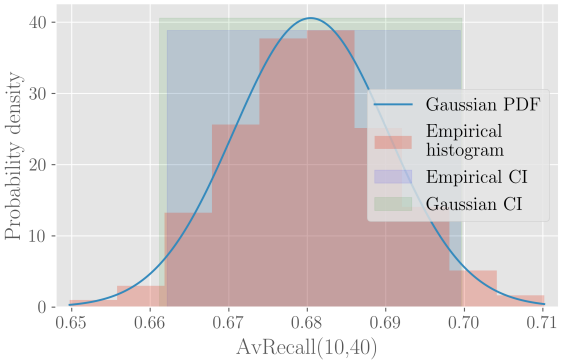


Figure 7: Empirical distribution with best-fit Gaussian for AvRecall(10,40) with 1000 bootstrap samples, with 95% uncertainty intervals. This is for the best model-encoding pair on the modeling split test set.

Table 18: Top-10 feature importance values with respect to Recall@20 using Permutation Feature Importance (PFI).

Feature	Meaning	Importance	Std. Dev.
BankDetailsAdded	This is a binary "yes/no" feature representing whether patients have their bank details added to Nikshay, signifying an increased likelihood of receiving monthly monetary government incentives for better nutrition during the TB treatment.	0.105	0.002
ReasonforTesting	Whether the test offered is for diagnosing TB or for assessing disease prognosis using follow-up testing.	0.060	0.002
CurrentFacilityPHIType	Binary variable: public/private. "Current" refers to the patient's preferred place of receiving medication and follow up. PHI (Peripheral Health Institute) is a health facility manned by at least one officer.	0.029	0.001
CurrentFacilityTBU	TBU is an administrative unit under the program with standardized coverage area, population and resources budgeted under the program.	0.026	0.001
CurrentFacilityDistrict	The district where a patient is seeking TB treatment and care, which may or may not be the place from where they originate get diagnosed.	0.017	0.001
CurrentFacilityPHI	Refer "CurrentFacilityPHIType". This refers to the actual PHI accessed by the patient.	0.015	0.002
resultSampleA	The specimen used for testing TB is received in two samples. This refers to the result for the first sample.	0.012	0.001
Age	Age of the patient.	0.009	0.002
TypeOfCase	Drug sensitive TB vs Drug resistant TB.	0.008	0.001
ContactTracing_Done	Whether the treatment supervisor visits the patient's home and assesses family contacts who are living in the same house. These contacts are at an increased risk of acquiring TB from them and are thus subject to mandatory screening for TB.	0.008	0.001

Appendix J. Improving Performance on Underperforming Cohorts

We explore two classes of methods to improve performance on underperforming cohorts: data augmentation and algorithmic fairness.

J.1. Data augmentation

We investigate a simple cohort-level data augmentation scheme that trains a separate model per cohort, trained on training data augmented with $N = 10$ copies of all patients from the category of interest. The idea here is to add extra signal in the training data corresponding to this category. When we apply this technique to districts, we find mean Recall@20 significantly improving for underperforming districts. For instance, the mean recall for all districts with Recall@20 < 0.3 changes from 0.175 to 0.328, a 87.43% increase, albeit with an increase in effective k. This is somewhat similar to increasing weight of patients from that district in the loss function. Our models are trained on the full mod-

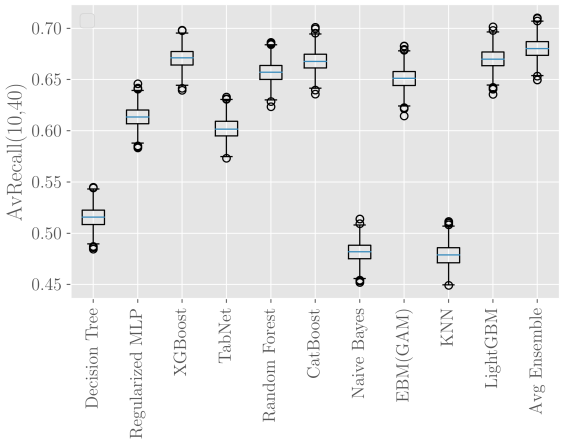


Figure 8: Boxplots with 1000 bootstrap samples for various models used with the similarity encoder, generated from the modeling split test set.

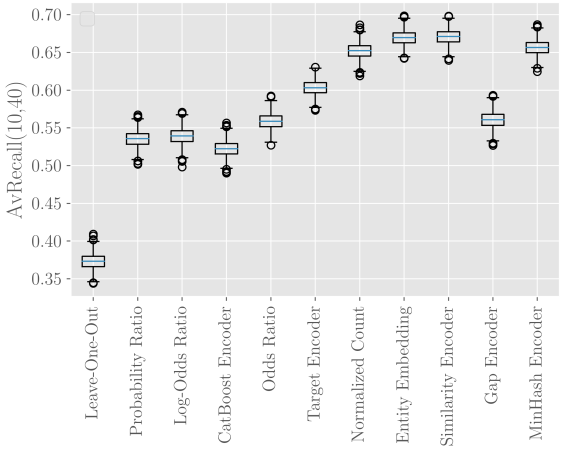


Figure 9: Boxplots with 1000 bootstrap samples for various encoders used with the XGBoost model, generated from the modeling split test set.

eling split, with results reported on the passive evaluation split. See (Table 19) for detailed results, reported using global thresholds.

Several models (logistic regression, LightGBM, XGBoost, random forest, and regularised MLP) have the option to use reweighted loss as a hyperparameter, but the augmentation method was found to do better on underperforming cohorts. We also investigated augmentation with log-inverse frequency (each patient is reweighed by $\log(N/N_d)$ where N is the total number of patients and N_d is the number of patients in that patient's district) but that does not yield any further benefit. Other data augmentation techniques such as SMOTE (Chawla et al., 2002) did not help either. This may be because performing augmentations with high-dimensional encoded categorical data may not be as useful as with other applications where interpolation is more natural.

J.2. Algorithmic fairness

For deployment in sensitive healthcare settings, it is critical that the model is fair across different important cohorts. When we analyse our model performance across cohorts to diagnose this issue, we note variance in performance across geographies, in particular districts, which is the unit of deployment. While the model does satisfactorily on non-location cohorts, when we look at gender, we note that it does worse on the female cohort.

To improve this, we explore a post-hoc fairness technique inspired by (Rodolfa et al., 2021) to equalize performance across these cohorts (Table 20 for states and Table 21 for districts and Table 22 for gender). The technique involves training a model, changing model scores of patients such that recall is roughly equalized across cohorts on a hold-out set, and then using the same model to evaluate on a test set, using these new cohort-level thresholds. The authors recommend these be forward-split in time. We train on the training set of the modeling split, and use global thresholds to compute Recall@20. We then use a combination of the test and val sets to tune cohort-level thresh-

Table 19: Results for data augmentation using $N = 10$ copies of the district of interest. The first column shows the max Recall@20 value for the districts aggregated for a particular row.

Threshold	Statistic	Recall@20$^{(\text{Original})}$	Recall@20$^{(\text{Mitigated})}$	# Districts	Lift%
0.2	Mean	0.080	0.400	4	402.832
	Std. dev.	0.093	0.437	4	369.975
0.3	Mean	0.175	0.328	9	87.316
	Std. dev.	0.108	0.294	9	172.990
0.4	Mean	0.296	0.355	27	20.155
	Std. dev.	0.108	0.175	27	61.993
0.5	Mean	0.404	0.437	73	8.170
	Std. dev.	0.109	0.145	73	33.100
0.6	Mean	0.472	0.485	133	2.638
	Std. dev.	0.112	0.129	133	15.088
1.0	Mean	0.586	0.575	222	-1.834
	Std. dev.	0.177	0.175	222	-1.428

olds, and evaluate on the passive evaluation split.

We get an increase in model performance across underperforming cohorts for all covariates analysed – states, districts and gender. We see all states but the best-performing show an improvement, and see particularly significant improvements for very poorly performing districts, with a 132% increase in mean recall for districts with Recall@20 $<$ 0.3. We only use districts that have at least one LFU patient in train, union of val and test, and passive evaluation sets, which reduces the total number of districts to 212 from 224. Our analysis for gender shows that application of the fairness technique reduces variation in performance between females and males. Overall results for all methods are summarized in Table 23.

Table 20: State-wise analysis of post-hoc threshold adjustment fairness technique, with recall and effective k before and after adjustment. We note an improvement for all states except the best performing state Uttar Pradesh.

State	Recall@20$^{(\text{Original})}$	Recall@20$^{(\text{Mitigated})}$	#LFUs	k$^{(\text{Original})}$	k$^{(\text{Mitigated})}$	Δ True +ves
Karnataka	0.518	0.750	156.0	11.838	27.140	36
Maharashtra	0.609	0.653	244.0	14.771	17.850	10
Uttar Pradesh	0.661	0.570	1108.0	24.986	18.744	-99
West Bengal	0.401	0.608	146.0	10.625	24.674	30

Table 21: Summary performance across districts using the algorithmic fairness technique, with recall and effective k before and after adjustment. We note a very significant increase in performance for underperforming cohorts.

Threshold	Statistic	Recall@20$^{(\text{Original})}$	Recall@20$^{(\text{Mitigated})}$	k$^{(\text{Original})}$	k$^{(\text{Mitigated})}$	# Districts
0.2	Mean	0.087	0.401	6.055	22.595	12
	Std. dev.	0.077	0.297	1.853	9.274	12
0.3	Mean	0.173	0.403	8.287	21.247	26
	Std. dev.	0.098	0.253	5.118	9.644	26
0.4	Mean	0.270	0.482	9.503	21.511	54
	Std. dev.	0.118	0.230	5.449	8.480	54
0.5	Mean	0.354	0.511	11.505	21.642	96
	Std. dev.	0.131	0.216	6.143	9.340	96
0.6	Mean	0.403	0.530	13.031	21.561	127
	Std. dev.	0.144	0.210	6.769	9.373	127
1.0	Mean	0.543	0.578	18.318	20.671	212
	Std. dev.	0.213	0.208	11.095	8.347	212

Table 22: Analysis of post-hoc threshold adjustment technique for the gender cohort, with recall and effective k before and after adjustment. We note a modest improvement in performance for the "Female" gender.

State	Recall@20$^{(\text{Original})}$	Recall@20$^{(\text{Mitigated})}$	#LFUs	k$^{(\text{Original})}$	k$^{(\text{Mitigated})}$	Δ True +ves
Female	0.560	0.589	582.0	16.751	18.754	16
Male	0.646	0.627	1071.0	22.332	20.894	-21
Transgender	0.750	0.750	1.0	22.963	20.741	0

Table 23: Summary of inequality measures across states, districts and gender upon application of the algorithmic fairness technique. We find improvements in mean and standard deviation of Recall@20 for states and gender, but a worsening of both for districts.

Metric	**Recall@20**$^{(\text{Original})}$	**Recall@20**$^{(\text{Mitigated})}$
Overall performance (State level)	0.616	0.606
Gini coefficient (State level)	0.099	0.057
Mean (State level)	0.547	0.646
Standard deviation (State level)	0.114	0.078
Overall performance (District level)	0.616	0.561
Gini coefficient (District level)	0.222	0.202
Mean (District level)	0.543	0.578
Standard deviation (District level)	0.213	0.208
Overall performance (Gender)	0.616	0.613
Gini coefficient (Gender)	0.065	0.055
Mean (Gender)	0.652	0.655
Standard deviation (Gender)	0.095	0.084