Unsupervised Risk Discovery for Child Welfare Screening

Sebastian Caldas [1] **Amanda Coston** [1] **Kin Guitierrez** [1] **Leqi Liu** [1]

Abstract

Child protection agencies across the US are routinely tasked with the screening millions of cases in order to identify those that are likely to result in adverse outcomes. To help with this task, many agencies have implemented and deployed risk prediction models that use administrative data in order to aid call workers. Nonetheless, the use of these predictive models for child welfare has proven onerous, raising concerns that go from how to validate their efficacy, to how to guarantee the fair treatment of disadvantaged groups. In particular, one concern that has inspired significant recent work is the fact that the variable of interest is never observed: we never see what happens to a child if the case was not investigated. In this work, we propose to tackle this obstacle by using unsupervised methods to discover latent risk structures in the agencies' administrative data, completely bypassing the "missing labels" predicament. We formulate two different risk models (with and without time dependencies among calls) and show (1) that the risk structures learned by both models are meaningful for experts, and (2) that taking into account time dependencies ends up hurting the performance on downstream predictive tasks.

1. Introduction

Machine learning algorithms have been used to improve and understand human decision in social contexts, and evidence suggests they can help to achieve better outcomes than unassisted human-decision makers (Kleinberg et al., 2017; Chouldechova et al., 2018). These algorithms are particularly helpful in settings where there is a wealth of information; too much for the human-decision maker alone to sift through. Such is the case for child welfare screening decisions in Allegheny County, Pennsylvania, where hot-line call screeners have to decide whether to "screen-in" a call for investigation based on the allegation and the county records for all parties associated with the allegation (i.e. child, parent, alleged perpetrator). Chouldechova et al. (2018) built a risk assessment tool to aid the call-screeners with this decision, and our work aims to upon their risk modeling by using unsupervised methods to discover latent risk structures in the child welfare data.

For child welfare screening, unlike standard classification problems, we do not observe the variable of interest. The variable of interest is the risk of harm to child under no investigation (i.e. screen-out) since assessment of this risk leads to optimal treatment assignment: If the child is safe from harm under screen-out, then there is no reason to spend county resources investigating the case. If the child is suffering or at risk of harm, then the county should open an investigation to determine how to mitigate this risk; such recourse may include offering services to support the family or in the extreme case, placing the child out-of-home. However, we have no labelled data on child harm under screen out. Prior work has used rereferral to the hotline at a later time as noisy proxy for child harm. This is noisy since there are reasons a call may be re-referred that do not reflect child harm. Prior work has also used placement out-of-home as a proxy for child harm under investigation. We note that the latter is not the same as the outcome of interest since the effect of investigation may mitigate the risk of harm.

We propose using unsupervised methods to discover latent risk structures for child welfare screening. We use Latent Dirichlet Allocation (LDA) to model latent risk types, and we extend LDA to include time dependencies in a dynamic topic model. We assign semantic meaning, such as *pattern of child abuse and substance abuse* to the topics learned in the standard and dynamic LDA models. The advantage of our approach is that we bypass the missing labels problem and we learn a feature representation that is interpretable by practitioners. The drawback is that we cannot fully validate our results. We use re-referral and placement outcomes to perform sanity checks for our models. While these outcomes are not sufficient to properly validate the model, we can use these models in conjunction with expert knowledge of the child welfare process to determine whether the model finds a meaningful representation of risk.

1.1. Project Objective

The objective of this work is to investigate whether an unsupervised learning approach can be used to understand

and improve the screening decisions, and help to overcome the limitations of supervised machine learning for this problem.

2. Related Methods

2.1. Child Welfare Risk Modeling

Chouldechova et al. (2018) trained a risk assesssment model for child welfare screening using placement as the outcome. This approach has the advantage of converting the risk problem into the classification setting where standard machine learning methods can be applied. The drawback of this supervised model is two-fold:

1. The outcome is only observed for cases that are screened-in (i.e. selective labels problem)
2. The model trains on a different potential outcome than the true outcome of interest.

To elaborate on the second drawback, we use the Neyman-Rubin potential outcome, denoting Y^a as the potential outcome Y that we would observe under intervention $A = a$ (Rubin, 2005). Letting A denote the screening decision where $A = 1$ corresponds to screen-in, then our outcome of interest, as motivated in the Introduction, is Y^0, whereas Chouldechova et al. (2018) are using Y^1 to train the model. If the intervention has any treatment effect, then $\mathbb{E}[Y^0|X] \neq \mathbb{E}[Y^1|X]$.

De-Arteaga et al. (2018) propose a reweighing procedure to resolve the selective labels problem which leverages "expert consistency" to resolve overlap violations. Inverse probability weighing (IPW) methods can be used to resolve selective labels if every person has a non-zero chance of being screened-in, which is referred to as overlap. De-Arteaga et al. (2018) propose imputing the label as 'no-harm' for people who have no overlap, justifying this procedure with the notion of expert-consistency.

However, no approach to child welfare screening has yet considered how to resolve limitation (2) and appropriately account for the effect of intervention on the observed outcome. This is a particularly notable limitation since if the child welfare process works as intended, the intervention should mitigate risk. The investigation may reduce risk because increased supervision may cause the perpetrator to change their behavior or because as part of the investigation the county may choose to offer services to aid the family that could include counseling, case management, or employment training. Conversely, in some cases, the investigation could increase risk of harm because the stress of increased supervision may exacerbate the situation. Therefore, we hypothesize that an unsupervised approach may discover a more realistic risk structure than a supervised approach which assumes no intervention effects.

2.2. Unsupervised Topic Modelling

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the most prevalent topic modeling method. For a given corpus, LDA learns: (1) the topics in the corpus and (2) the topic distribution for each document. A significant advantage of LDA is that it does not require labels of the documents. Possible methods for evaluating LDA include interpreting the semantic meaning of the topics and performing downstream tasks on the documents like clustering (and see if the clusters are meaningful). In our setting, we treat *children* as *documents*. By ignoring the labelling information, e.g. the screen-in and screen-out decision for each child, LDA not only ensures the learned topics to be semantically meaningful, but guarantees the topic distribution for each child to be unbiased.

2.3. Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBN's) are temporal directed acyclic graphical models. Hidden Markov Models (HMM's) and Kalman Filters are two common DBNs; DBNs generalize these structures to include any directed acyclic graph with repeating units over time, where the units are referred to as *time slices* (Kanazawa et al., 1995). A topic model can be formulated as a DBN e.g. (Blei & Lafferty, 2006). This dynamic topic model allows both the distribution over topics to change over time, as well as the definition of topics (i.e. the distribution over words). In this setting, variational inference is used because of the nonconjugacy that arises in the dynamic formulation which models time dynamics as Gaussian.

3. Method

3.1. LDA for Static Risk Modeling

We build a static model that uses Latent Dirichlet Allocation (LDA) to uncover any underlying structure in the child records. Each child record can be seen as a document c that has N_c features. The whole dataset is then treated as a corpus with C child records. Given this corpus, LDA will provide us with topics $\{\beta_k\}_{k=1}^K$, i.e. representative distributions over the features $\mathcal{F}$ that have semantic meaning. Each child record c is assigned a distribution over the topics. Then, a screen-in decision for the child can be made by checking if c has higher probability of falling in the high risk topic.

To use LDA, we make two simplifying assumptions on the dataset: (1) features of the same call are independent from each other; (2) call records are independent, i.e. call records about the same child at different time are treated as different documents. The graphical model of the child welfare data is shown in Figure 1. The generative process of the child welfare data set is described as follows:

- For $k = 1, ..., K$:
  - Draw topic $\beta_k \sim \text{Dirichlet}(\eta)$.
- For every call record $c \in \mathcal{C}$:
  - Draw per-call-record mixture proportion $\theta_c \sim \text{Dirichlet}(\alpha)$
  - For each feature f in the record description c:
    - Sample a topic indicator $z_{cf} \sim \text{Multi}(\theta_c)$
    - Sample the feature $w_{cf} \sim \text{Multi}(\beta_{z_{cf}})$

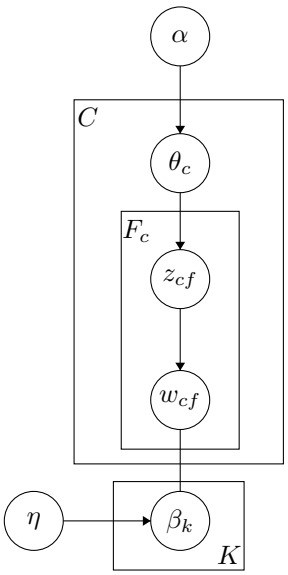

Figure 1. Graphical Representation of the Latent Dirichlet Allocation on the Child Welfare Dataset.

Inference We use the standard gensim (Řehůřek & Sojka, 2010) library for the LDA implementation. As suggested by the library, the inference method is batch Variational Bayesian inference (Hoffman et al., 2010).

Evaluation LDA is in general hard to evaluate since there is no ground truth of topics in a corpus. We perform two types of evaluation:

1. Interpret the semantic meanings of the discovered topics
2. Predict downstream outcomes using the per-child-record mixture as an embedding of the child record. These downstream outcomes are A) whether the child will be re-referred to the hotline in a six month period from the original referral and B) whether the child will be placed out-of-home.

We conduct the downstream predictions using two approaches: 1) *observational evaluation:* the standard approach that uses observed outcomes and 1) *counterfactual evaluation*: a counterfactual approach that controls for the effect of interventions on the observed outcome. In particular, the county's investigation and any services offered can mitigate risk, so observationally someone who received this treatment looks "lower risk" than they would have been counterfactually had they not received treatment. We believe our counterfactual analysis is more accurate since it accounts for intervention effects; we present the results of both methods in Section 4.

The counterfactual evaluation relies on three standard assumptions from causal inference, where A denotes the decision to investigate and where $\pi(X) = \mathbb{P}(A = 1|X)$ denotes the propensity score:

1. Consistency: $Y = AY^1 + (1 - A)Y^0$.
   This assumes there is no interference between treated and untreated units. This is a reasonable assumption in the child welfare screening setting since it is unlikely that opening an investigation into one case will affect another case's observed outcome.
2. Exchangeability: $Y^0 \perp A|X$. This assumes that we measured all variables X that jointly influence the intervention decision A and the potential outcome Y^0.
3. Weak positivity requirement: $\mathbb{P}(\pi(x) < 1) = 1$ requires that each example have some non-zero chance of being screened out.

With these assumptions we can identify counterfactual error metrics; the derivation for precision is given below and the other error metrics are described in the Appendix.

Letting $\hat{h}$ denote the predicted label, the target counterfactual precision is

$$\mathbb{E}[Y^0 \mid \hat{h} = 1]$$

Under our causal assumptions this is identified as

$$\mathbb{E}[\mathbb{E}[Y \mid X, A = 0] \mid \hat{h}(X) = 1]$$

Letting $\mathbb{P}_n(f)$ denote the sample average of f, the doubly robust estimator for counterfactual precision is

$$\mathbb{P}_n\Big[\frac{1 - A}{1 - \hat{\pi}(X)}(Y - \hat{\mu}_0(X)) + \hat{\mu}_0(X) \mid \hat{h}(X) = 1\Big]$$

3.2. Dynamic Topic Models for Sequential Risk Modeling

To relax the assumption that phone calls are time independent, we use a dynamic topic model to capture the time dependencies among the phone calls. Adapting the dynamic topic model proposed by (Blei & Lafferty, 2006), we keep

the topics over time fixed and consider the topic mixture for each phone call to be time dependent. Such a design ensures the risk profiles across time to be comparable. As shown in Figure 2, we assume the topics (risk types) remain constant over time; as time changes, the risk profiles (the topic mixtures) will be drawn from a different distribution. The generative process is described below (Blei & Lafferty, 2006):

- At time slice t, draw $\alpha_t, \alpha_{t-1} \sim \mathcal{N}(\alpha_t|\delta^2 I)$.
- For each phone call $i \in \mathcal{I}$ at time slice t:
  - Draw $\eta_{ti} \sim \mathcal{N}(\alpha_t, \sigma^2 I)$.
  - For each $k \in [K]$, $\theta_{tik} = \frac{\exp(\eta_{tik})}{\sum_k \exp(\eta_{tik})}$.
  - For each feature:
    - Sample a topic $z_{tif} \sim$ Multi(θ_{ti}).
    - Sample a feature $w_{tif} \sim$ Mult$(\beta_{k,z_{tif}})$.

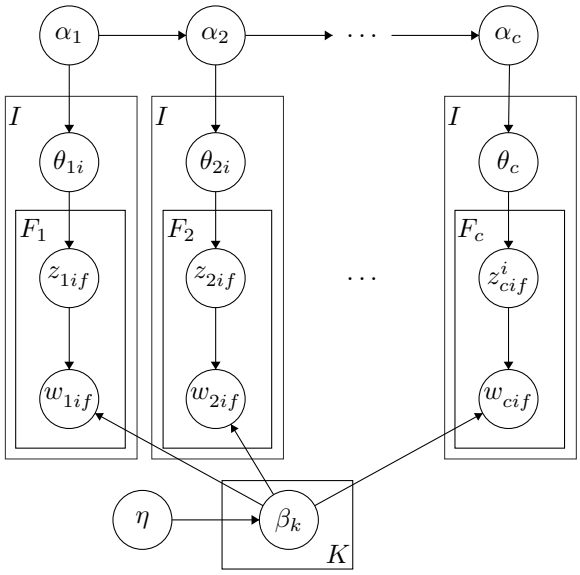

Figure 2. Graphical Representation of the Dynamic Topic Model on the Child Welfare Dataset.

Since the child welfare process is inherently temporal, our goal here is to investigate whether a model that allows time dependencies learns more meaningful risk profiles that might perform better on downstream prediction tasks.

Inference Since this dynamic model does not have conjugate prior, as described in (Blei & Lafferty, 2006), a more ideal inference method is variational inference with a mean-filed approximation distribution. The approximate variational posterior is

$$\prod_{k=1}^{K} q(\beta_k|a_k) \times \prod_{c=1}^{C} \left(\prod_{i=1}^{I} q(\theta_{ti}|b_{ti}) \prod_{f=1}^{F_t} q(z_{tif}|c_{tif}) \right).$$

The update rules are then derived based on variational kalman filtering and variational wavelet regression.

Evaluation We hypothesize that the phone call number is a confounder of the child risk profiles. For example, a second phone call of the same child implies that the child is at a higher risk. Alternatively, the first phone call record could have a positive or negative effect on the child. To test this hypothesis, we explicitly model this confounder through the previous dynamic topic model and build a classifier conditioned on the phone call time slice. We compare the performance of the dynamic model versus the static model conditioned on the time slice of the phone call.

4. Experiments

In this section we present experimental results for the models presented in Section 3.1 and Section 3.2.

4.1. Data

We use data from Allegheny County's Department of Human Services (ACDHS). The ACDHS data consists of 58,468 calls to the hotline between 2010 and 2016 and includes county records, screening decisions, and coded information from the phone call. These calls are over 30,000 in total; 48% of these calls were screened in for investigation, and 13% of these screened-in cases resulted in a placement outcome (Chouldechova et al., 2018). The features in the county records include demographic and socioeconomic information, historical child welfare interaction, public welfare, usage of public programs like Supplemental Security Income and Temporary Assistance for Needy Families, involvement with the Allegheny County criminal justice system, and behavioral health information for Medicaid recipients. This data is part of ACDHS's Data Warehouse, which links data across all publicly funded human services. The call log data codes the alleged child maltreatment into categories including inadequate clothing, left alone, sexual abuse or exploitation.

4.2. Data Preprocessing

One obstacle of applying LDA to the child welfare dataset is translating a child record into a document. There are four kinds of features among the feature set, namely binary, continuous, ordinal and categorical. Binary features are only included if their value is `True` for the given child. For categorical and ordinal features, we treat their values as words. For each feature f^i that takes values $\{f^i_j\}_{j=1}^d$, a child that

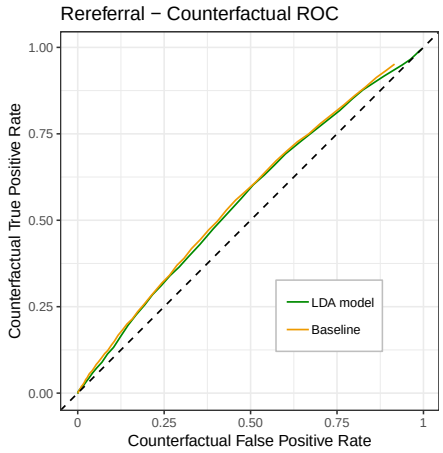

Figure 3. Counterfactual ROC curve for re-referral task

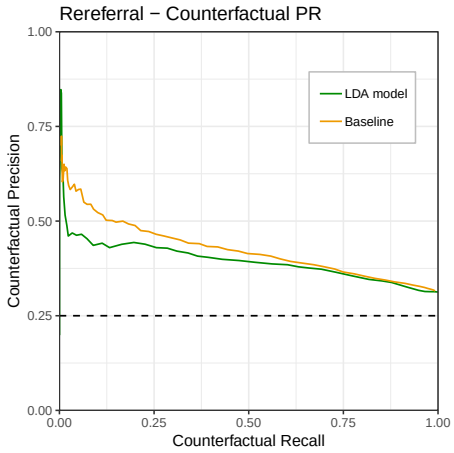

Figure 4. Counterfactual PR curve for re-referral task

has feature $f^i = f^i_j$ will have the word "$f^i : j$" in his/her document. For continuous features, we first put the values into 10 bins and rank them as if the features are ordinal. Then, we perform the same processing for the continuous features as the one we have done for the ordinal features.

4.3. LDA for Static Risk Modeling

We trained an LDA model according to Section 3.1, using the code provided by Řehůřek & Sojka (2010). We set the number of topics $K = 15$, a number that allows us to have a reasonable set of interpretable risk profiles that still hold predictive power. All other hyperparameters, except for the number of passes (which was set to 100) were left as those specified by default.

4.3.1. Semantic Analysis

We assigned semantic meaning to the 15 topics learned by the static LDA model (see Table 1). These meanings were assigned by analyzing the features that were most probable given the topic. We observe a mix of expected and unexpected topics: for instance, topics 10 and 12 are expected since drug/alcohol abuse, domestic violence, and malnutrition or lack of proper supervision are standard allegations of child abuse/neglect. Topics such as topic 5 *homeless, young child* and topic 6 *low risk, previous referral* are more surprising. Future work could validate the semantic meaning of these topics by comparing to call screeners' assessments of the most common risk profiles.

4.3.2. Performance on Downstream Tasks

We define two downstream tasks in order to verify that our learned risk profiles do capture information about each call's latent risk. Specifically, we use each call's distribution over risk profiles as its feature representation, and use it to predict two relevant risk proxies: (1) whether a given record will be re-referred in the future, and (2) whether any subsequent investigation results in placement out-of-home. For each one of these tasks, we train a random forest using 5-way cross validation. We tune the number of trees and the maximum depth, leaving all others as the defaults in (Pedregosa et al., 2011). We use the same train-test splits as those in (Chouldechova et al., 2018).

We obtain the ROC and Precision-Recall (PR) curves shown in Figure 6 using the standard observational approach. We also plot a counterfactual ROC curve and PR curve using our counterfactual estimates of true positive rate, false positive rate, and precision (see Figures 3 and 4). Both methods of evaluations (standard and counterfactual) show that our LDA representation performs only slightly worse than using the full feature set (retaining $\sim 95\%$ of the ROC AUC and $\sim 82\%$ of the PR AUC). Thus, we conclude that the learned risk profiles capture the most salient information about the call's latent risk.

4.4. LDA for Dynamic Risk Modeling

A dynamic topic model is trained as described in Section 3.2 using the code provided by Řehůřek & Sojka (2010) and setting the number of topics to 15 (to allow for a fair comparison with our static model). Our main hypothesis to test is whether conditioning on the phone call will improve the classifier performance. Throughout the section, we consider 5 different time slices corresponding to the first four calls for a given case plus a time slice for calls corresponding to fifth calls and beyond. In other words, each call that presents a new case is assigned to time slice t_1; each call that re-refers a case for the first time is assigned to t_2; second re-referrals are assigned t_3; third to t_4, and all other calls go into t_5.

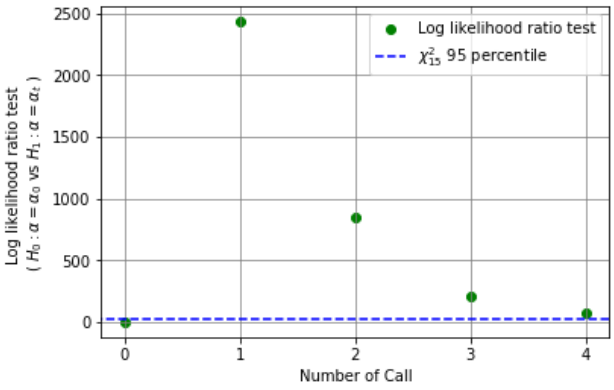

Figure 5. Likelihood ratio test for the contrast between the fixed topics model and the dynamical topics model. Number of call 0 refers to the static LDA model.

4.4.1. Dynamic and Fixed Topics Test

We first test whether the data suggests that risk profiles vary over time. To do this, we test for the goodness of fit of the risk profiles of our sequential model vs. those of our static model. We use the likelihood ratio test in which we compare the null hypothesis that the generative process of the risk profiles has a common Dirichlet distribution (like in the classic, static LDA), against the alternative hypothesis that the risk profiles come from time varying distributions (like in our sequential model). Now, since it is not possible to obtain a closed expression for the distribution of the statistical test, we rely on the Wilkinson's asymptotic convergence of the ratio test over the variational approximations of the original likelihoods. As we can see from Figure 5 there exist significant evidence that suggests the existence of dynamical risk profiles.

4.4.2. Semantic Analysis

We display our interpretation of the given topics in Table 2. Topic 3, long-term pattern of inadequate physical care, is the most common topic, which makes sense since inadequate physical care is also the most common allegation to the child welfare hotline, comprising 39% of calls. There is some overlap with the topics from the static model in Table 1, such as topic 4 and 9, but there are notable differences. Particularly, we see very specific topics, such as topics 6 and 7, *chronic medical neglect* and *sexual assault with low risk of repeat incident*.

4.4.3. Performance on Downstream Tasks

Finally, knowing that our dynamic and static risk profiles are statistically distinct (see Section 4.4.1), we examine the performance of the risk profiles learned through our dynamic model on the same two tasks specified in Section 4.3.2. More specifically, we wish to test whether conditioning on the phone call time slice will improve the classifier performance. To do this, for each time slice t: (1) we train a model on calls assigned to t, and (2) we test the model on a held-out set of calls also assigned to t. For each model, we follow the same training procedure described in Section 4.3.2.

We are interested in how these different models perform when using the risk profiles extracted from the dynamic topic model as features. As our baseline, we train each model on the risk profiles extracted from the static LDA presented in Section 3.1. We present the results in Table 3 (for the re-referral task) and Table 4 (for the placement out-of-home task). It is clear that the features learned by the dynamic topic model are not as predictive of our risk proxies as those obtained from the static LDA. Furthermore, the performance of the conditioned models (i.e., trained and tested on calls from one time-slice) is comparable to their performance on a global model (i.e., trained on calls from all time-slices). We are led to conclude that the temporal information captured by our dynamic topic model is not only insufficient to help in these downstream tasks, but that the proposed conditioning is hurtful for them.

	Semantic interpretation
Topic 1	Inadequate parental care.
Topic 2	Uncertain risk, referral history.
Topic 3	Criminal history and public welfare assistance programs.
Topic 4	High risk infant, parent with criminal history.
Topic 5	Homeless, young child.
Topic 6	Low risk, previous referral.
Topic 7	Referral history, economic assistance.
Topic 8	Pattern of child abuse and substance abuse.
Topic 9	Low risk inadequate care by mother.
Topic 10	Drug and alcohol use, domestic violence.
Topic 11	High risk, pattern of abuse.
Topic 12	Malnutrition and/or lack of supervision of toddler or young child.
Topic 13	Teen victim with prior child welfare history and criminal involvement.
Topic 14	High risk of maternal neglect.
Topic 15	Inadequate physical care, impending danger.

Table 1. Semantic Interpretation of the Topics obtained through Static LDA.

	Semantic interpretation
Topic 1	Unstable home situation and economically impoverished.
Topic 2	High risk and impending danger
Topic 3	Long term pattern of inadequate physical care.
Topic 4	Public welfare assistance and criminal involvement.
Topic 5	Criminal history and referral history.
Topic 6	Chronic medical neglect.
Topic 7	Sexual assault with low risk of repeat incident.
Topic 8	Impending danger to newborn/ infant.
Topic 9	High risk, pattern of abuse.
Topic 10	Criminal history of mother.
Topic 11	Criminal history of alleged perpetrator, school age child.
Topic 12	Referral from other county or agency.
Topic 13	Inadequate food, public welfare assistance.
Topic 14	Prior abuse or neglect history.
Topic 15	Low risk, mother has history of referrals.

Table 2. Semantic Interpretation of the Topics obtained through Dynamic LDA.

Time Slice	Static LDA	Dynamic Topic Modeling
1st phone call	ROC = 0.60 PR = 0.24	ROC = 0.50 PR = 0.18
2nd phone call	ROC = 0.63 PR = 0.30	ROC = 0.51 PR = 0.21
3rd phone call	ROC = 0.62 PR = 0.40	ROC = 0.50 PR = 0.31
4th phone call	ROC = 0.40 PR = 0.43	ROC = 0.50 PR = 0.35
5th or more phone calls	ROC = 0.64 PR = 0.61	ROC = 0.53 PR = 0.52

Table 3. Re-referral performance comparison of the Static LDA features and the Dynamical Topic Model.

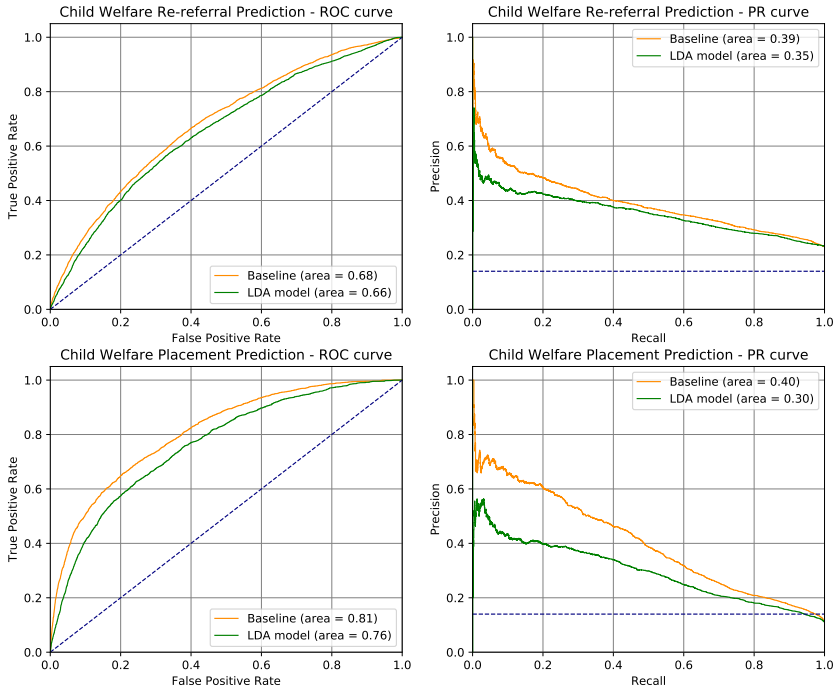

Figure 6. ROC and PR curves for our downstream tasks using the standard observational approach.

Time Slice	Static LDA	Dynamic Topic Modeling
1st phone call	ROC = 0.77 PR = 0.27	ROC = 0.52 PR = 0.09
2nd phone call	ROC = 0.77 PR = 0.28	ROC = 0.51 PR = 0.10
3rd phone call	ROC = 0.71 PR = 0.30	ROC = 0.50 PR = 0.16
4th phone call	ROC = 0.72 PR = 0.38	ROC = 0.53 PR = 0.18
5th or more phone calls	ROC = 0.69 PR = 0.33	ROC = 0.50 PR = 0.16

Table 4. Placement performance comparison of the Static LDA features and the Dynamical Topic Model.

5. Conclusions

We presented two unsupervised models that abstract the problem of child welfare screening in Allegheny County, Pennsylvania. Inspired by the widely used LDA model, our proposed solutions bypass the selective labeling problem encountered by previous work for this case study (Chouldechova et al., 2018; De-Arteaga et al., 2018) while perhaps requiring more domain knowledge in order to interpret the topics generated by the LDA (which we can interpret as different populations of children with varying degrees and causes of risk).

Our experiments showed that the LDA topic features perform almost on par with the raw features on downstream risk tasks in the child welfare process, and we observed that while the risk profiles do appear to change over time, allowing a model to capture these dependencies did not improve performance on the downstream risk tasks. Future work should focus on investigating whether allowing dependencies between the features improves performance on the placement and re-referral tasks.

References

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 134–148, New York, NY, USA, 23–24 Feb 2018. PMLR. URL http://proceedings.mlr.press/v81/chouldechova18a.html.

De-Arteaga, M., Dubrawski, A., and Chouldechova, A. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905*, 2018.

Hoffman, M., Bach, F. R., and Blei, D. M. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

Kanazawa, K., Koller, D., and Russell, S. Stochastic simulation algorithms for dynamic probabilistic networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 346–351. Morgan Kaufmann Publishers Inc., 1995.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1): 237–293, 08 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx032. URL https://dx.doi.org/10.1093/qje/qjx032.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

A. Appendix

A.1. Counterfactual error metrics

Let t denote the threshold used for classification, and let $\hat{h}(X) = \mathbb{I}\{\hat{\mu}(X) \geq t\}$ be the predicted label, where $\mathbb{I}$ is the indicator function. The target counterfactual FNR is

$$\mathbb{E}[1 - \hat{h}(X) \mid Y^0 = 1]$$

Using our causal assumptions in Section 3, this is identified as

$$\frac{\mathbb{E}\big[(1 - \hat{h}(X))\mathbb{E}[Y \mid X, A = 0]\big]}{\mathbb{E}\Big[\mathbb{E}[Y \mid X, A = 0]\Big]} \tag{1}$$

where we can use doubly robust estimates for the two iterated expectation terms.

The doubly robust estimate for the numerator is

$$\mathbb{P}_n\left[(1-\hat{h}(X))\Big[\frac{1-A}{1-\hat{\pi}(X)}(Y-\hat{\mu}_0(X))+\hat{\mu}_0(X)\Big]\right] \tag{2}$$

The doubly robust estimate for the denominator is

$$\mathbb{P}_n\Big[\frac{1-A}{1-\hat{\pi}(x)}(Y-\hat{\mu}_0(X))+\hat{\mu}_0(X)\Big] \tag{3}$$

Since recall/sensitivity is $1 - FNR$, we can use this estimator to also estimate recall.

A.2. False Positive Rate (FPR) Estimator

The target causal FPR is

$$\mathbb{E}[\hat{h}(X) \mid Y^0 = 0]$$

Using our causal assumptions in Section 3, this is identified as

$$\frac{\mathbb{E}\Big[\hat{h}(X)\mathbb{E}[1 - Y \mid X, A = 0]\Big]}{\mathbb{E}\Big[\mathbb{E}[1 - Y \mid X, A = 0]\Big]} \tag{4}$$

where as the doubly robust estimate for the numerator is

$$\mathbb{P}_n\left[\hat{h}\Big[\frac{1-A}{1-\hat{\pi}(X)}(\hat{\mu}_0(X)-Y)+(1-\hat{\mu}_0(X))\Big]\right] \tag{5}$$

The doubly robust estimate for the denominator is

$$\mathbb{P}_n\Big[\frac{1-A}{1-\hat{\pi}(x)}(\hat{\mu}_0(X)-Y)+(1-\hat{\mu}_0(X))\Big] \tag{6}$$