Early Alert Systems During a Pandemic: A Simulation Study on the Impact of Concept Drift

Yang Xu
yang_xu@brown.edu
The Policy Lab at Brown Univeristy

Kevin H. Wilson
kevin_wilson@brown.edu
The Policy Lab at Brown University

ABSTRACT

Predictions from early alert systems are increasingly being used by institutions to assist decision-making and support at-risk individuals. Concept drifts caused by the 2020 SARS-CoV-2 pandemic are threatening the performance and usefulness of the machine learning models that power these systems. In this paper, we present an analytical framework that uses imputation-based simulations to perform preliminary evaluation on the extent to which data quality and availability issues impact the performance of machine learning models. Guided by this framework, we studied how these issues would impact the performance of the high school dropout prediction model implemented in the Early Warning System (EWS). Results show that despite the disruptions, this model can still be reasonably useful in assisting decision-making. We discuss the implications of these findings in more general educational contexts and recommend steps in countering the challenges of using predictions from imperfect machine learning models in early alert systems and, more broadly, learning analytic research that uses longitudinal data.

CCS CONCEPTS

• **Applied computing** → **Education**; **Computing in government**; Computer-assisted instruction.

KEYWORDS

early alert system, concept drift, dropout prediction, imputation, simulation, pandemic

ACM Reference Format:
Yang Xu and Kevin H. Wilson. 2021. Early Alert Systems During a Pandemic: A Simulation Study on the Impact of Concept Drift. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3448139.3448190

1 INTRODUCTION

More institutions are adopting early alert systems powered by risk prediction models to assist decision-making [9]. Trained with historical data, these machine learning models uncover patterns in the relationships between certain risky outcomes (e.g., dropping out of school) and behavioral and performance metrics that tend to be associated with them. The models then produce "risk scores" that indicate the likelihood of these outcomes for individuals in the future based on past behavior and/or performance and flag those whose risk scores are above certain thresholds. This prompts decision-makers to take actions in supporting the individuals that the models deem at-risk.

The 2020 SARS-CoV-2 pandemic has highlighted the need for more discussions on the impact of concept drifts in early alert systems research in the learning analytics community. The term "concept drift" refers to the phenomenon where changes in data over time cause decay in the performance of predictive models [12]. The stream learning literature has systematically studied this issue [4, 12], focusing on developing efficient online algorithms for automatic drift detection [6] and adaptive model updating [3, 5, 13] in systems that process large volumes of streaming data. In the context of early alert systems, the pandemic has caused disruptions to data collection practices that will likely alter the quality and potentially the meaning of education data gathered during its course. It also may substantively alter people's behavior, potentially rendering the machine learning models trained on pre-pandemic data much less useful. Therefore, we need principled ways to think about how to make predictive models useful despite the impact of concept drifts. First, we need to evaluate the extent to which concept drift degrades model performance. Second, we need to consider the sociopolitical implications of using results produced by less-than-optimal models, especially given the increasing scrutiny that model-assisted decision-making is under for concerns over its fairness and transparency [8, 10, 15].

In this study, we propose an analytical framework to evaluate the impact of the SARS-CoV-2 pandemic on machine learning models in early alert systems trained with pre-pandemic historical data. We focus on using a real-world example of a high school student dropout prediction model implemented in the Early Warning System (EWS) piloted by the Rhode Island Department of Education (RIDE) in collaboration with The Policy Lab at Brown University and DataSpark. Guided by the definition of concept drift from stream learning literature [12], we enumerate several potential patterns in pandemic-induced data quality issues and perform several imputation-driven simulation studies of their potential effects using available historical data. Our goal is to understand as best as we can how risk models' predictions may be impacted by the pandemic. In particular, we explore:

- How, if at all, can the existing EWS model still be useful in making dropout risk predictions on the School Year (SY) 2019-20 data impacted by pandemic-related disruptions?

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448190

- Assuming that schools reopen and in-person instruction resumes early in SY 2020-21, how, if at all, can future EWS models incorporate such impacted data?
- Assuming that schools remain closed for at least a large proportion of SY 2020-21, how can future EWS models account for this "new normal?"

To answer these questions, we open with a brief background on the EWS and situate the challenges of concept drift in both data quality and in the current sociopolitical context. We will then describe in detail the existing dropout prediction model and introduce the analytical framework built on related work on concept drift that addresses these challenges. After presenting the results from the simulation study, we will discuss the usefulness of these results and the political implications of using imputations in this and more general contexts.

2 BACKGROUND AND RELATED WORK

2.1 The Early Warning System

Rhode Island's Early Warning System is designed to assist administrators and teachers in identifying 9th through 12th Grade students who are at risk of dropping out and connecting those students with resources they need to be successful. The system features a teacher information portal, which displays the latest available data on six student performance indicators updated monthly: attendance percentages, grade retention, suspensions, math proficiency, English Language Arts (ELA) proficiency, and a risk indicator of dropping out. The risk indicator is color-coded with red, yellow, or green to indicate high, moderate, and low risk of dropping out at the beginning of each school year.

The risk model—including what data to include and how to display risk bands—was developed collaboratively by RIDE, The Policy Lab, and DataSpark. It is produced by a machine learning model trained on historical student outcome data on all public high school students in Rhode Island from between SY 2007[1] and 2015. We will furnish more details on the model itself in Section 3. For now, we emphasize that although the model is trained with dropping out as the outcome, the goal of the risk indicator is to, in addition to the five other performance indicators, provide the teachers and administrators another way to identify students most in need of help. For this purpose, EWS provides educators with links to resources most relevant to the students at risk of dropping out.

After a period of user testing and small-scale releases, the EWS was successfully released to all teachers and administrators in February 2020, one month before the SARS-CoV-2 pandemic closed down schools and instruction started moving online. Due to uncertainty in model performance under changed circumstances, RIDE decided to temporarily disable the predictive model but retain the other performance indicators on their dashboard. Before relaunching the model, RIDE asked The Policy Lab to explore how several potential challenges to the EWS's underlying data might impact model performance, In particular, the widespread shutdowns and online instruction might cause changes in definitions in attendance and grievances. In addition, Rhode Island cancelled standardized testing for SY 2019 [2], so some standardized test scores will not be available to the model.

We also need to consider the sociopolitical implications of using less accurate predictions. A recent example from the UK where a model was used to impute test scores for standardized tests cancelled due to the pandemic make this particularly salient [1]. Although the EWS risk indicator was not designed to assign accountability or otherwise punish or promote, there could still be stigma associated with being identified as "at-risk." Perhaps most importantly, we need to communicate transparently and effectively both the usefulness and the caveats of using such imperfect predictions.

2.2 Concept Drift

Concept drift is the problem that how and whether data is collected and how it relates to outcomes changes over time. It is an issue in machine learning that had received much scholarly interest before the pandemic because it happens to almost all machine learning models that rely on longitudinal data. There have been many attempts over the past decade to precisely define concept drift and chart the landscape of all efforts to address this issue especially in streaming machine learning literature [4, 11, 12, 16]. Among these efforts, [12] performed a comprehensive review of the concept drift literature and offered a most succinct definition. In the remainder of this paper, we rely on their definition as follows:

Let X_t denote the vector of features collected at time t and Y_t denotes the label corresponding to each record in X_t, then concept drift occurs when one or both of the following is observed:

(1) the distribution of data changes: $P(X_{t+1}) \neq P(X_t)$;
(2) the underlying relationship between the features and labels changes: $P(Y_{t+1} \mid X_{t+1}) \neq P(Y_t \mid X_t)$.

In this paper, we will be concerned primarily with the first situation. While it is possible (perhaps even likely) that the pandemic has radically altered the relationship between the EWS's features and its labels (whether a student will drop out), we will be unable to validate that assumption until well after SY 2020 ends. A policymaker who believes that this relationship in a predictive model has radically changed should consider suspending the use of the model until more data can be gathered.

3 METHODS

We begin this section by furnishing more technical details on the EWS risk prediction model. Then we will use the definition of concept drift to construct the analytical framework that we use to evaluate the impact of the pandemic under different scenarios.

3.1 The EWS Model

The model that powers the dropout risk indicator in EWS is a random forest [7] model. RIDE uses a .NET-based system, so we trained the model using the `FastForest` implementation in Microsoft's `nimbusml` package for Python and incorporated the model into RIDE's student information system. The training set contained historical year-end outcome data on all Rhode Island public school students from 9th to 11th Grade between 2007 and 2015. Features include:

[1] For simplicity, school years are referenced by the year of the Fall semester.

- Demographic information: gender, age, race, ethnicity, socioeconomic statuses, and special education statuses;
- Attendance and grade retention: rate of attendance and whether the student repeated a grade;
- Grievances: numbers of infractions and of days a student was suspended both in- and out-of-school;
- Test scores: scores from state standardized tests, and SATs and PSATs if available.

The labels of the training set were whether a student eventually dropped out in or before 12th Grade. We used data from the two previous years to make predictions when possible. We also applied techniques such as a combination of over- and under-sampling to counter the issue of imbalance and down-weighting less-recent records from the same individual to achieve better predicative performance. We tested the model performance on the test set which consists of identically structured data from 2016.

Before productionizing the model, The Policy Lab, DataSpark, and RIDE collaboratively made several modeling decisions. For instance, in one meeting, a longtime RIDE employee mentioned that RIDE's suspension policy had dramatically changed in SY 2012, leading to a discussion of how to weight infraction data. We also collaboratively examined the model's performance on subgroups within the test set to ensure equitable performance for each subgroup as well as to determine where the model would not be helpful. For instance, Rhode Island's juvenile detention facility has a dramatically different dropout rate than other schools in the state, and its students have dramatically different needs than others. So a decision was made both to exclude it from the training data and to not deploy it to that school.

Table 2 shows the how the model performed on each of the metrics that we used. In addition to the frequently used metrics to evaluate machine learning models such as overall accuracy, precision, recall, F1 scores, and AUC scores, we also used three metrics specific to this context: the predicative accuracy among the top 100, 3%, and 5% individuals ranked by the risk scores produced by the model. This is because in practice, the EWS will mark the top 3% of the individuals with the highest risk scores as at "high" risk of dropping out, and the subsequent 2% as at "moderate" risk of dropping out. These cutoffs were chosen by RIDE with input from many of their domain experts. They considered in particular: 1) the capacity to perform interventions, 2) the perceived "utility" of the rating, 3) the actual dropout rate, 4) technical limitations, and 5) the ethical impacts of over/under-labeling students.

3.2 Analytical Framework

This analytical framework focuses on the problem that the pandemic causes sufficient data quality issues that corrupt at least some features. Formally, at a certain time $t+1$ after the onset of the pandemic, we obtain a dataset X_{t+1} with at least some corrupted or missing features due to disruptions in data generation or collection. We have an existing model M_t trained with a dataset X_t and labels y_t collected before the disruptions. We are concerned with the extent to which M_t can still produce reasonably accurate predictions $\hat{y}_{t+1}$ from X_{t+1}, because not all features in X_{t+1} are equally useful in making the predictions. If $\hat{y}_{t+1}$ is reasonably reliable, these predictions can still have some practical value, whereas M_t should no longer be used if $\hat{y}_{t+1}$ produces nonsensical results. However, we do not have true labels y_{t+1} associated with X_{t+1} for us to make such evaluations.

In this paper, we simulate this issue through imputation. Since we know that some features in X_{t+1} will be corrupted or missing, we will likely need to impute the values for these corrupted features. We can simulate this workflow using past data. Specifically, we set aside a subset of X_t and the corresponding labels y_t, which we will call X_s and y_s. We can then treat some features $x_i \in X_s$ as if they were corrupt or missing and instead impute these features to produce a synthetic dataset X_{imp}. The dataset $\{X_{imp}, y_s\}$ can be used in conjunction with the rest of $\{X_t, y_t\}$ to simulate different scenarios in which we make predictions.

This framework requires a strong assumption to be meaningful. Although there are two possible sources of concept drift the definition in Section 2.2, simulations will only address source (1), changes in data distributions over time. If source (2)—changes in the relationship between features and labels—also contributes substantially to the drift, it is likely that $P(y_s \mid X_{imp}) \neq P(Y_t \mid X_t)$, which means we cannot use $\{X_{imp}, y_s\}$ to approximate $\{X_{t+1}, y_{t+1}\}$. Other methods such as [14] may be able to provide worst-case bounds for the model's performance in such a situation, though they require their own assumptions on how much the distribution $P(Y_{t+1} \mid X_{t+1})$ has changed from $P(Y_{t+1} \mid X_{t+1})$.

We will work under this assumption for the rest of this paper. This is not to deny that source (2) may contribute to drift, and in future work we hope to explore ways to measure its effects. In the end, policymakers must make a decision as to whether an early alert system's utility outweighs its potential downsides. Indeed, leaving an early alert system trained on historic data unchanged implicitly assumes that either the early alert system remains more useful than harmful or that the early alert system will have little effect at all on actual educator and student behavior. Our work provides a quick and inexpensive way for researchers and practitioners to check the robustness of their existing model against data quality issues. We hope that by highlighting the complex practical and ethical issues raised by the pandemic and contributes to discussions among the learning analytics community on how to proactively address its effects on the ever-growing number of early alert systems deployed in the field.

3.3 The Three Scenarios

Below we outline three scenarios we will explore with simulation studies. They are graphically represented in Figure 1.

3.3.1 Scenario 1: Prediction during concept drift. In this scenario we evaluate the effectiveness of M_t given new unlabeled data X_{t+1} subject to concept drift. In the context of EWS, this scenario focuses on the most immediate questions of whether we can use the same EWS model trained on data from SY 2019 and before to predict risks of dropping out in SY 2020, knowing that SY 2019 data is most likely to be unreliable with missing fields such as infractions and suspensions. These metrics might not be well defined with schools switching to remote learning for the second half of the school year. Meanwhile, test scores are also missing due to cancellations of standardized tests.

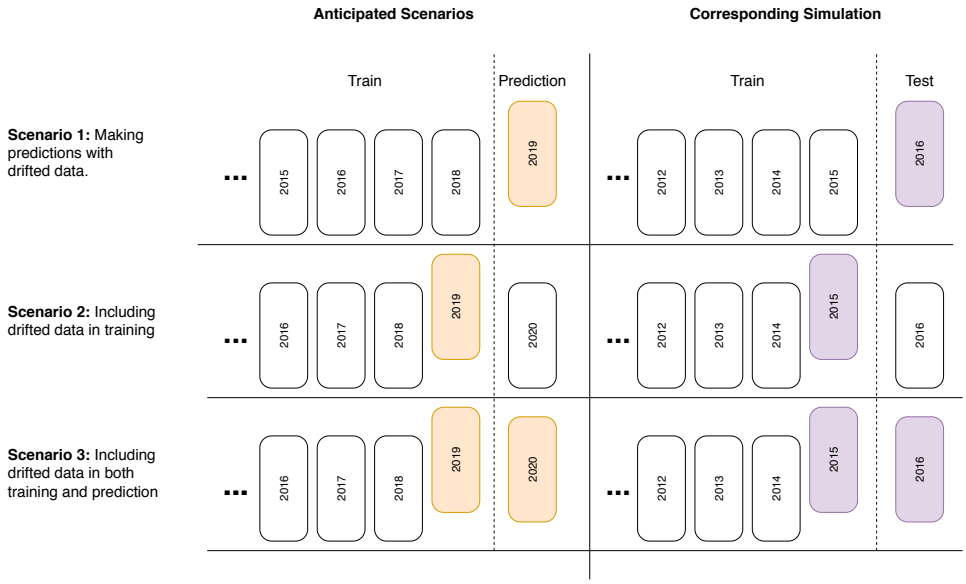

Figure 1: Anticipated Scenarios vs. the corresponding simulation strategies.

To simulate this scenario, we dropped the true values from the test set and imputed new ones to create a synthetic dataset X_{imp}. We then fed the synthetic dataset to the same model M_t to see whether the model could still produce accurate predictions on imputed data. We used different imputation strategies for different features. For standardized testing scores, we used cumulative maximum scores for each individual in the place of their actual scores in the test set. For attendance rates and infractions and suspensions, we used the imputation strategies listed in Figure 1.

3.3.2 Scenario 2. Normalcy returns. In this scenario we evaluate the case where the pandemic ends and we assume that new data collected at this point is distributed similarly data collected before the pandemic. In other words, we have collected a new dataset X_{t+2} after the abrupt drift subsides, which is similarly distributed to X_t collected before the pandemic. We want to assess whether a new model M_{t+1} trained on $\{X_{t+1}, y_{t+1}\}$ combined with $\{X_t, y_t\}$ can be used in the future. In this and the following scenario, we are concerned with model performance immediately after the pandemic subsides and before we have collected new data.

For the EWS, this and the next scenario explore beyond SY 2020. This scenario focuses on the situation where the disruption from the pandemic is temporary. We are tasked with making predictions on "normal" data with models trained on historical data that contain years affected by SARS-CoV-2. We simulate this condition as follows: first, we train new models on modified training sets. To produce these sets, we use the same imputation strategies mentioned in the previous scenario on the year immediately before the year in the test set. Then, we re-train the models on these modified training sets and evaluate the models on the test sets.

3.3.3 Scenario 3. A "new normal" is established. In this scenario, we evaluate the case where the pandemic has fundamentally modified people's behavior so that the data never returns to the same distribution as before the pandemic. In other words, we have collected a new dataset X_{t+2} some time after the abrupt drift subsides, which is differently distributed than X_t collected before the pandemic and X_{t+1} collected during the pandemic. We want to evaluate whether a new model M_{t+1} trained on $\{X_{t+1}, y_{t+1}\}$ combined with $\{X_t, y_t\}$ can be used for better performance in the future.

For EWS, this scenario is similar to the previous one with a key difference: we do not assume that things will go back to normal. To simulate this condition, we apply the same imputation strategy used on the training sets in the previous scenario to the test set as well.

3.4 Checking model performance and equity

In addition to comparing performance metrics with the original model, in applicable cases, it is imperative that we also consider the model's differential performance within subgroups. This is especially tricky as data quality issues caused by the pandemic may impact different subgroups differently. However, reporting these metrics and interrogating their plausibility with policymakers is critical to the success of any early alert system.

4 RESULTS

In this section we report the simulation results produced from different imputation strategies on all of the metrics that we used to evaluate the original model. We reiterate that in each of the scenarios, we used cumulative maximum scores from all other years for standardized test scores. The imputation strategies reported only refers to how we imputed attendance rates, numbers of infractions, and number of days in in-school and out-of-school suspensions.

Table 1: Imputation strategies

Strategy	Description
Set to Zero	Use 0 for infraction and suspension fields and keep attendance intact
Mean	Use mean of all other years to fill attendance, infraction, and suspension fields
Median	Use median of all other years to fill attendance, infraction, and suspension fields
Nearest neighbor	Use values from previous year to fill attendance, infraction, and suspension fields
Worst	Use smallest attendance rate and largest numbers of infraction and suspension from all other years to fill corresponding fields

4.1 Scenario 1: Can the existing model still be used?

For this scenario we test the original model using the data from the last year replaced with imputed data. Table 2 suggests that with most imputation strategies, the model performance does not deviate much from the performance achieved on the original unmodified test set. If we set infractions and suspensions to 0 while not changing the attendance rates, the model performance stays virtually the same. We postulate two reasons that might be behind this surprising result. First, the numbers of infractions and suspensions are not as "important" as attendance rates during prediction, which is consistent with the feature importance scores produced by the model during training. Second, the model looks back at the performance of the previous year as well, which was intact. This practice lends some robustness to the model's response to sudden changes in one year.[2]

4.2 Scenario 2: What if things return to normal?

For this scenario, we impute on one year's data within the training set and use this training set to train new models. Because of the way our data is structured, the imputed values will also propagate to the corresponding "last year" fields in the test set. Results (Table 3) show that model performance was again not significantly impacted by using imputation on one year's training data. Tables 2 and 3 report very similar results. Again, we postulate that this is because the model looks at two years' performance at any given time, which adds some robustness to sudden changes in data.

4.3 Scenario 3: What if there is a "new normal"?

The setup for this scenario is similar to the previous one except that we also introduce the imputation used in Scenario 1 to the test set. In other words, both the values for the current year and the previous year in the test set are now imputed. We do not expect that the model performance would deviate too much from what we have seen in Tables 2 and 3. Table 4 shows that it is indeed the case. However, the last column of Table 4 is particularly interesting. When tested with the worst performance from each student as imputed values, the model has increased recall yet decreased precision, while the overall performance seems to be on par with the other models. This indicates that the model now has more false positives than false negatives. We speculate that since neither the "current year" nor the "previous year" performance data is accurate under this scenario, some of the robustness that we discussed previously is now lost. This is especially more apparent when we compare the last columns of Tables 3 and 4.

[2]We note that this, in and of itself, is an interesting result as it suggests that earlier interventions (say in 9th Grade) based primarily off of attendance triggers might have large effects on eventual dropout.

5 DISCUSSION

5.1 The utility of the results

We first discuss the utility of the results beyond what is presented in the previous section. At first glance, our study might seem to be context-specific and our results context-dependent. Indeed, the finding that our model might still produce reasonable predictive performance on imputed data could be a function of the data itself and the way we structured the data and built the model. However, we argue that the challenges of concept drift that we faced are universal to early alert systems. Our study is a first attempt to quantify the effects of the pandemic on early alert systems built using longitudinal data, a problem that should be addressed with some urgency.

Urgency is the keyword here. Our simulation framework offers one method to explore the potential degradation in model performance caused by the pandemic. In particular, it does not rely on new data, which will not be available until many months after the model's predictions might be useful. Indeed, while this study was underway, we simultaneously requested more recent data from partners, a process that itself takes time. Still, the preliminary results presented here facilitated our partners' short-term planning for the EWS for SY 2020. In that sense, we believe that our framework does offer some structure in thinking about the present and the future of existing systems.

In addition, we learned from our simulations that these predictive models are more robust to abrupt disruptions might be helpful to counter the immediate impact of the pandemic. Our practice of looking at more than one year's data seemed to be in play here. Future research should examine what practices increase early alert systems' robustness.

Finally, recall that our explorations relied on the strong assumption that the pandemic has not radically altered the relationship between features and labels. This is perhaps an optimistic assumption. However, if it is not true, then policymakers should consider whether their uncertainty in a model's performance outweighs the model's potential utility. In future work, we plan to explore how we might quantify this uncertainty, though [14] and related literature offers some potential avenues.

Table 2: Simulation results under Scenario 1

Metric	Original Model	Set to Zero	Mean	Median	Nearest Neighbor	Worst
Overall Accuracy	0.9130	0.9120	0.9138	0.9133	0.9120	0.9074
NPV	0.9502	0.9488	0.9538	0.9550	0.9499	0.9423
Precision	0.4812	0.4757	0.4847	0.4811	0.4755	0.4541
Recall	0.5057	0.5086	0.4753	0.4562	0.4973	0.5258
F1 Score	0.4931	0.4916	0.4799	0.4683	0.4861	0.4874
AUC	0.8902	0.8840	0.8791	0.8765	0.8798	0.8785
Accuracy @ Top 3%	0.6452	0.6462	0.6278	0.6278	0.6350	0.6431
Accuracy @ Top 5%	0.5622	0.5628	0.5536	0.5518	0.5567	0.5671
Accuracy @ Top 100	0.83	0.83	0.81	0.81	0.82	0.77
N Positive Examples in Test Set	2731	2731	2731	2731	2731	2731

Table 3: Simulation results under Scenario 2

Metric	Original Model	Set to Zero	Mean	Median	Nearest Neighbor	Worst
Overall Accuracy	0.9130	0.9126	0.9127	0.9151	0.9109	0.9018
NPV	0.9502	0.9482	0.9503	0.9572	0.9491	0.9316
Precision	0.4812	0.4795	0.4797	0.4921	0.4697	0.4346
Recall	0.5057	0.5225	0.5020	0.4537	0.4921	0.5756
F1 Score	0.4931	0.5001	0.4906	0.4721	0.4804	0.4953
AUC	0.8902	0.8910	0.8907	0.8851	0.8855	0.8882
Accuracy @ Top 3%	0.6452	0.6524	0.6370	0.6237	0.6350	0.6513
Accuracy @ Top 5%	0.5622	0.5788	0.5653	0.5622	0.5586	0.5727
Accuracy @ Top 100	0.83	0.85	0.78	0.83	0.85	0.80
N Positive Examples in Test Set	2731	2731	2731	2731	2731	2731

Table 4: Simulation results under Scenario 3

Metric	Original Model	Set to Zero	Mean	Median	Nearest Neighbor	Worst
Overall Accuracy	0.9130	0.9115	0.9122	0.9147	0.9097	0.8954
NPV	0.9502	0.9469	0.9526	0.9610	0.9503	0.9227
Precision	0.4812	0.4741	0.4754	0.4886	0.4610	0.4135
Recall	0.5057	0.5240	0.4698	0.4083	0.4650	0.5965
F1 Score	0.4931	0.4978	0.4726	0.4448	0.4630	0.4885
AUC	0.8902	0.8856	0.8792	0.8683	0.8743	0.8833
Accuracy @ Top 3%	0.6452	0.6503	0.6094	0.6002	0.6104	0.6319
Accuracy @ Top 5%	0.5622	0.5757	0.5536	0.5450	0.5377	0.5665
Accuracy @ Top 100	0.83	0.85	0.80	0.83	0.84	0.83
N Positive Examples in Test Set	2731	2731	2731	2731	2731	2731

5.2 The sociopolitical implication of imperfect predictions

Perhaps more important is the discussion of the sociopolitical implications of using imperfect predictions from early alert systems. Our study shows that the EWS model could still be useful even though the data might not be perfect, which lends confidence to stakeholders to continue incorporating the model in the EWS. One recurring theme in our discussion with stakeholders, however, is the cost of wrong predictions. While no models are perfect and wrong predictions are bound to occur, the fact that the model is imperfect and more likely to make mistakes could potentially outweigh its usefulness. Whether to continue using an imperfect model is a highly contextual decision to be made, but we offer the following recommendations in working with stakeholders on such decisions.

- **Lower the stakes.** This will lower the cost of wrong predictions and redirect focus on the usefulness of the predictions on assisting decision-making considering all other factors.
- **Ensure transparency.** Be candid in communicating the limitations of model predictions to both decision makers and the public. Be very specific about why the models can be less than optimal but still useful.

- **Ensure fairness.** Checks are needed to ensure that no subgroup is impacted more negatively than others. These checks should already be in place to guard against inherent bias in data but are more relevant now to prevent any techniques used to adjust models against concept drifts from putting some subgroups at a disadvantage.

6 CONCLUSION

In this paper, we used concept drift to conceptualize the challenges that the 2020 SARS-CoV-2 pandemic brings to early alert systems research and learning analytics research in general that uses longitudinal data. Our simulation framework offers a time-efficient way to produce preliminary evaluation on the impact of concept drift on data and predicative modeling and outlines ways to respond to the immediate and short-term difficulties caused by the pandemic. We showed that while imperfect, some predictive models can still be useful under certain circumstances. However, we also highlight the sociopolitical implications of continuing to use such models. Such decisions are to be made cautiously with stakeholders and extra steps need to be in place to ensure the transparency and fairness of these models.

ACKNOWLEDGMENTS

The authors are funded by Arnold Ventures. We thank Ariel Neumann, Melanie Bowdish, Kimberly Bernard, David Yokum, Ben Guhin Delphine, Matthew Santacroce, Attiyya Houston, and the anonymous reviewers for support and feedback on this work. We especially thank Jennifer LoPiccolo, Elizabeth Texeria, Elizabeth Laudry, Fidel Achille, and Sophie Tan of RIDE for their leadership in the EWS project.

REFERENCES

[1] BBC. 2020. A-level and GCSE results: Pressure mounts on ministers to solve exam crisis. https://www.bbc.com/news/education-53804323

[2] Linda Borg. 2020. RICAS tests cancelled because of coronavirus. *The Providence Journal* (2020). https://www.providencejournal.com/news/20200403/ricas-tests-canceled-because-of-coronavirus

[3] Luis Eduardo Boiko Ferreira, Luiz S Oliveira, Heitor Murilo Gomes, and Albert Bifet. 2019. Adaptive Random Forests with Resampling for Imbalanced data Streams. (2019), 6.

[4] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4 (April 2014), 1–37.

[5] Heitor M. Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. 2017. Adaptive random forests for evolving data stream classification. *Mach Learn* 106, 9-10 (Oct. 2017), 1469–1495.

[6] Manzoor Ahmed Hashmani, Syed Muslim Jameel, Vali Uddin, and Syed Sajjad Hussain Rizvi. 2020. Concept Drift Detection Technique using Supervised and Unsupervised Learning for Big Data Streams. (2020), 13.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Random forests. In *The elements of statistical learning*. Springer, 587–604.

[8] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proc. of the CHI Conf. on Human Factors in Comp. Sys. 2019* (2019), 1–16.

[9] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauría, James R. Regan, and Joshua D. Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Learning Analytics* 1, 1 (May 2014), 6–47.

[10] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Binz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 45–55.

[11] Imen Khamassi, M. Sayed Mouchaweh, Moez Hammami, and Khaled Ghédira. 2018. Discussion and review on evolving data streams and concept drift adapting. *Evolving Systems* 9 (2018).

[12] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under Concept Drift: A Review. *IEEE Trans. Knowl. Data Eng.* (2018), 2346–2363.

[13] Jacob Montiel, Rory Mitchell, Eibe Frank, Bernhard Pfahringer, Talel Abdessalem, and Albert Bifet. 2020. Adaptive XGBoost for Evolving Data Streams. In *Proc International Joint Conference on Neural Networks (Glasgow, UK)*. IEEE.

[14] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. 2020. Evaluating Model Robustness to Dataset Shift. *arXiv:2010.15100 [cs, stat]* (Oct. 2020).

[15] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proc. of the CHI Conf. on Human Factors in Comp. Sys. 2018 - CHI '18* (2018), 1–14.

[16] Indrė Žliobaitė. 2010. Learning under Concept Drift: an Overview. *arXiv:1010.4784 [cs]* (Oct. 2010).