A Longitudinal Evaluation of a Deployed Predictive Model of Fire Risk

Jessica Lee
School of Computer Science
Carnegie Mellon University
jl5@andrew.cmu.edu

Yanwen Lin
College of Engineering
Carnegie Mellon University
yanwenl@andrew.cmu.edu

Michael A. Madaio
School of Computer Science
Carnegie Mellon University
mmadaio@cs.cmu.edu

Abstract

As machine learning systems become increasingly deployed to augment civic decision-making and inform public policy, it is critical that we evaluate the stability and robustness of their performance. In this paper, we evaluate a predictive model of fire risk deployed in a mid-sized American city to inform the prioritization of commercial property fire inspections. We describe here our evaluations of the model's performance over the 7 months since its deployment, including analyses of the stability of the risk scores assigned to properties and the calibration of the model's performance for commercial properties in urban areas above and below the poverty line. Given the high-stakes involved as machine learning systems are deployed in civic life, regular, transparent evaluations such as these are essential in informing citizens and stakeholders of the efficacy of predictive models.

1 Introduction

Building fires pose a serious risk to the lives and livelihoods of residents of urban centers around the world, with nearly 475,000 building fires in the US in 2016 alone, causing over 3,000 civilian deaths, and over $10 billion in property damage [2]. As the UN reports that the world's population is increasingly living in urban areas [16], urban fire departments are increasingly asked to engage in fire risk reduction efforts like property inspections and fire safety education, [2] while grappling with ever more limited resources. This suggests an opportunity for predictive models to help fire departments better prioritize their risk reduction efforts. Our prior work [17] developed and deployed a predictive model of the likelihood of a fire incident at a given address in a 1-year window, using an XGBoost model [6] trained on 8 years of data on fire incidents (and other fire calls from the fire department), as well as property-level features (e.g. property assessments, sale price, lot area, etc), and deployed this model with the City of Pittsburgh's Bureau of Fire (see Figure 4 for the model pipeline). This work built off of prior work in fire risk modeling, such as Firebird [15] in Atlanta and RBIS [9] in New York City, by incorporating temporal features such as property code violations and other fire department calls (e.g. smoke alarms, gas leaks, etc), and by deploying the risk model with the city's fire bureau. However, given the high stakes involved, and the risk that machine learning models may become "stale" if the input data changes (among other reasons) [11], we report here on our approach to a longitudinal evaluation of the performance and impact of our deployed fire risk model over the first 7 months of its deployment, inspired by others' work on evaluations of deployed models for welfare risk prediction, student dropout risk, and recommendations for machine learning models used in public policy, among others [1, 8, 14]. We discuss here our approach to conducting a suite of performance evaluations and impact assessments of our deployed commercial fire risk model. While we also developed a fire risk model for residential properties, aggregated at the census block (i.e. several hundred addresses) to inform community fire safety education, this model has not yet been deployed, and thus, we report here primarily on evaluations of the commercial risk model, with a discussion of a simulated approach to evaluating the residential model included in the appendix.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

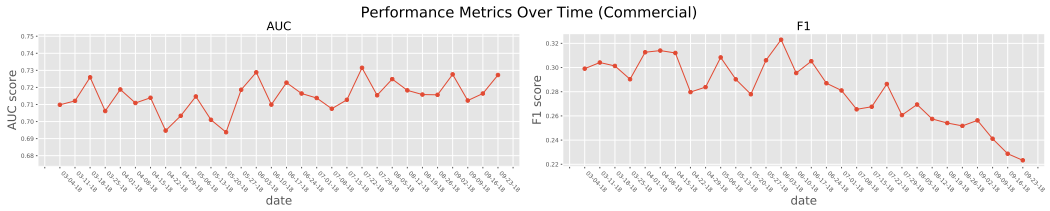

Figure 1: Raw values of AUC and F1 for model iterations from March 4th to September 23rd

2 Performance Evaluation

2.1 Model Performance

In order for fire personnel to most effectively use the results of our model to augment their inspection decision-making, it is critical that our model continues to rank the most risky properties highly over time as new data is obtained. We thus report in this paper on the raw values, means, and standard deviation of a set of model performance measures (i.e. accuracy, AUC, recall, precision, F1, and kappa) from the logs of the retrained commercial model from its first iteration on March 4th, 2018 to September 23rd, 2018. In Figure 1, we can see that, while the AUC has remained relatively stable over the deployment, the F1 metric has decreased from an initial 0.30 to 0.22. As F1 is the harmonic mean of precision and recall, we can see in the Appendix that, while the recall of the model has improved somewhat throughout the deployment (from 0.46 to 0.48), the precision of our model has decreased over time (from 0.21 to 0.14), which explains the corresponding decrease in F1. However, in consultation with fire department officials at the time of deployment, they expressed a preference for a more conservative approach to identifying risky properties, favoring recall (e.g. minimizing false negatives) over precision (e.g. minimizing false positives), and leading us to tune our initial hyper-parameters to optimize for greater recall. Further, as the model retrained on a weekly basis, the hyper-parameters were not re-tuned at each iteration. To understand how this performance differs at different risk levels, we computed the precision and recall at the top-k properties (ordered by risk level), and find that a smaller set of the most risky properties may have a more optimal balance between precision and recall than the full set of properties (see Figure 8 in the appendix).

2.2 Feature Importance Consistency

As the model used in the deployed system is a tree-based model (XGBoost), it splits the data on specific feature values, and so we wanted to understand changes in these features as the model retrains over time. We report here on changes in the feature importance from March 4th, 2018 to September 23th, 2018 using a heat map (Figure 9 in the appendix) to display the change of the top 10 features week by week. In Figure 9, the darker colors represent greater feature importance, for each weekly model iteration. The most predictive feature for several early model iterations is alarm system activation (Fire Code 745), while in later model iterations the most important features are Fire Code 743 and 5001, both related to smoke alarms. We also calculated the standard deviation of the feature importance score over the deployment to understand which features might be less consistently predictive. We can see in Figure 10, that Fire Code 745 (alarm system activation) has the most variance in its feature importance over time. This suggests further work on feature engineering to account for these variances in feature importance among the smoke detector and alarm codes.

2.3 Risk Score Stability

In our prior work [17], we describe how we discretized the risk scores (i.e. prediction probabilities) into three risk categories through discussion with the fire inspectors and operations chiefs using the system (low risk = 1-3, medium risk = 4-6, and high risk = 7-10). Because the risk scores are generated anew for each property each week as the model retrains, we assessed the stability of each property's risk score across each weekly model iteration, inspired, in part, by Ackermann et al.'s analysis of the *list stability* of officers at risk of adverse incidents [1]. We wanted to understand, as new risk scores are generated for each property in each weekly model retraining, the extent to which the proportion of properties in each risk category may change, as well as how likely it is that any

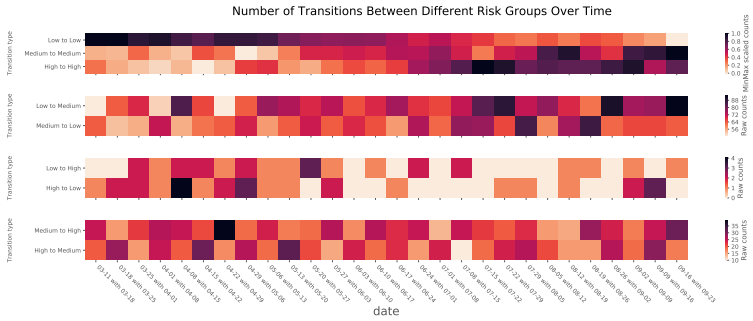

Figure 2: Number of properties transitioning between risk categories

Table 1: Actual Fire Incidents and Inspection Violations after Model Deployment

Risk Category	Total Properties	Code 111's	Code 111-118	Code 100-199s	Any Fire Code	Inspection Violations	No Violations	Avg. Violations/Property
7-10	89	8 (9.0%)	26 (29%)	33 (37%)	39 (44%)	12 (13%)	2 (2.2%)	3.4
4-6	325	3 (1.0%)	17 (5.2%)	29 (8.9%)	51 (16%)	80 (25%)	6 (1.8%)	4.1
1-3	12962	35 (0.27%)	99 (0.76%)	139 (1.1%)	382 (2.9%)	279 (2.2%)	32 (0.25%)	4.0

given property will have a sufficient change in its predictive probability to transition from a low risk category to a medium or high risk category, or vice versa. We first analyzed the percent of properties in each risk bracket at each model iteration from March 4th, 2018 to September 23rd, 2018. We find that there is a 1% decrease in the proportion of properties rated as low risk over the deployment, and a small increase in the proportion of properties rated as medium risk (from 0.035 to 0.042) – see Figure 11 for more detail. In Figure 2, we show the number of properties that transitioned across risk categories (e.g. low to low, low to medium, low to high, etc) in each successive model iteration. For properties that remained in the same risk category, we normalized it by the min and max values of the row due to the greater number of properties in the low risk category. We see that there are infrequent transitions from high to low and low to high (max of 4 properties), although there is an increase in the number of properties transitioning from low to medium risk.

2.4 Post-Hoc Prediction Performance

To evaluate how well our model actually predicted the properties at risk of fires, we identify the number and percent of fire calls of various types at properties in each of the risk categories since the model was deployed, as another way of evaluating the model's precision. These include Code 111s (building fires); Code 111-118s (any structure fires, including confined cooking fires, etc.); Code 100-199s (any fire-related incident, including trash and vegetation fires); and any other Fire Code logged after model deployment (e.g. smoke alarm activation, gas leak, etc). We also identified the number of inspections and the rate of inspection violations found in properties at each risk level. In Table 1, we see there is a greater percentage of Code 111s, Code 111-118s, Code 100-199s, and All Fire Codes in the high risk category than in medium risk, and a greater percentage in the medium risk category than low risk, which suggests that our model may be identifying the riskiest properties. In the 6-month period after the commercial risk model was deployed, 29% of the high risk properties had some type of structure fire incident (Code 111-118), compared to 5.2% of the medium risk properties and 0.7% of the low risk properties. Of the 14 high risk properties that were inspected for fire safety, violations were found in 12 properties, although this rate was similar for other risk categories. Future work should empirically investigate the relationship between fire safety inspections, violations, and the occurrence of fire incidents.

2.5 Differential Predictive Accuracy

Due to the significant examples of deployed machine learning systems having disparate impact on particularly vulnerable populations ([3, 7] to name only a few), we investigated the fire risk model's differential predictive accuracy (what [8, 13] refer to as calibration) for commercial properties located within lower-income neighborhoods compared to those in higher-income neighborhoods. Prior work on socio-economic factors associated with fires suggests that low-income communities may be at

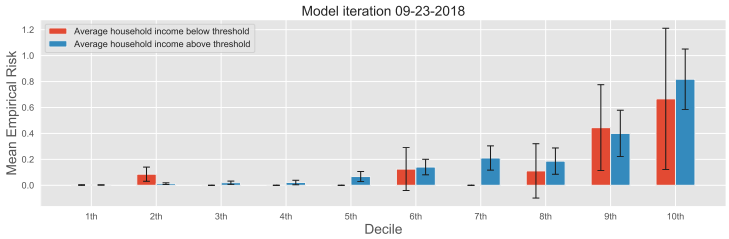

Figure 3: Observed fire incident rates for commercial properties in census blocks with weighted average household income below and above the poverty line. Error bars are 95% confidence interval.

greater risk of fire [12], but such communities may also have issues with the quality and coverage of their property data [10]. According to the U.S. Census Bureau, poverty thresholds are calculated at a household level, based on the income and the number of children and age of children in the household [4]. However, because the granularity of the available household income data from the American Community Survey is reported at a census block-level (comprised of hundreds of households), we thus calculated an average weighted poverty threshold for each census block [5]. We used the poverty income threshold of a k-person household, weighted by the percent of k-person households in the census block, averaged across all k values in the census (1 to "7 or more"), and separated commercial properties into those in census blocks with the weighted average household income above or below the poverty threshold. For both groups (above and below the threshold), we calculated the observed fire incident rate at each risk score decile (i.e. the ceiling of the predicted risk probability) following [8, 14]. As Chouldechova et al. describe, a model is well-calibrated if each decile is monotonically increasing and the proportion of items in each decile has an equivalent rate of adverse events across populations [8]. We find evidence of poor calibration at lower risk deciles for commercial properties in census blocks below the poverty line. In Figure 3, we see that commercial properties in census blocks below the poverty line at the 2nd risk decile have an equivalent rate of fire incident as commercial properties in blocks *above* the poverty line at the 5th risk decile. That is, at low risk levels, the fire risk of commercial properties in low-income census blocks is underestimated. We are planning to investigate why this might be the case, and how to mitigate it, as well as other measures of potential disparate impact [8, 7, 13].

3 Discussion

As machine learning models are deployed to augment civic decision-making and influence public policy, it is critical that we continue to monitor and evaluate the robustness of these models over time. As others have suggested [1, 8], and which we echo here, deployed predictive models may require criteria or measures of their effectiveness beyond the typical accuracy metrics often reported for machine learning models. In this paper, we report on results from a suite of such evaluations of a deployed predictive model of fire risk. We find that although our model does continue to successfully identify risky properties in which building fire incidents occur, the precision of our model has decreased over time, suggesting the need to re-tune hyper-parameters more regularly and the potential benefit of limiting the number of properties recommended for inspection to a top-k list. Additionally, further work is needed to better understand the likelihood for any given property to transition from one risk category to another (e.g. low to high), and investigate potential causal factors for these transition probabilities. Finally, we find evidence for some miscalibration of the fire risk model's performance for commercial properties in low-income census blocks at low risk levels, suggesting future work to understand and mitigate the source of this miscalibration. Future work should also investigate more sophisticated model types that may account for the potential spatial and temporal auto-correlation of incidents, as well as online or reinforcement learning methods for a deployed model to improve over time. We intend for this work to contribute to the growing body of work on transparently monitoring and evaluating the performance, stability, and robustness of deployed machine learning models, particularly models used to inform public policy and civic decision-making. We hope others may learn from our findings here to develop and improve on our models of fire risk to contribute to greater public safety and community resilience.

4 Acknowledgments

We thank our partners at Pittsburgh Bureau of Fire for being willing to examine and improve their processes, and we thank our partners at the Department of Innovation and Performance for their assistance with data acquisition and model deployment. We are indebted to many at the Metro21 Smart Cities Initiative for the stimulating conversation and financial support provided by the Hillman Foundation. Finally, we thank many others for conversations which improved this work, from Dr. Hinds-Aldrich at the NFPA, Shira Mitchell, and the other members of the Atlanta Firebird team.

References

[1] Klaus Ackermann, Joe Walsh, Adolfo De Unánue, Hareem Naveed, Andrea Navarrete Rivera, Sun-Joo Lee, Jason Bennett, Michael Defoe, Crystal Cody, and Lauren Haynes. 2018. Deploying Machine Learning Models for Public Policy: A Framework. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 15–22.

[2] National Fire Protection Association. 2017. (October 2017). `http://www.nfpa.org/News-and-Research/Fire-statistics-and-reports/Fire-statistics/Fires-in-the-US`

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[4] U.S. Census Bureau. 2018. Poverty Thresholds. `https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html`. (September 2018). Accessed on 10-5-18.

[5] United States Census Bureau. 2016. 2016 ACS 5-year estimates. `https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml`. (2016). Accessed on 10-5-18.

[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[8] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[9] Eddie Copeland. 2015. BIG DATA IN THE BIG APPLE. *Capital City Foundation.* (2015). `http://capitalcityfoundation.london/wp-content/uploads/2015/06/Big-Data-in-the-Big-Apple.pdf`

[10] Ben Green, Alejandra Caro, Matthew Conway, Robert Manduca, Tom Plagge, and Abby Miller. 2015. Mining administrative data to spur urban revitalization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1829–1838.

[11] Rishabh Iyer, Nimit Acharya, Tanuja Bompada, Denis Charles, and Eren Manavoglu. 2018. A Unified Batch Online Learning Framework for Click Prediction. *arXiv preprint arXiv:1809.04673* (2018).

[12] Charles R. Jennings. 2013. Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature. *Fire Safety Journal* 62, PART A (2013), 13–19. `DOI:http://dx.doi.org/10.1016/j.firesaf.2013.07.002`

[13] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

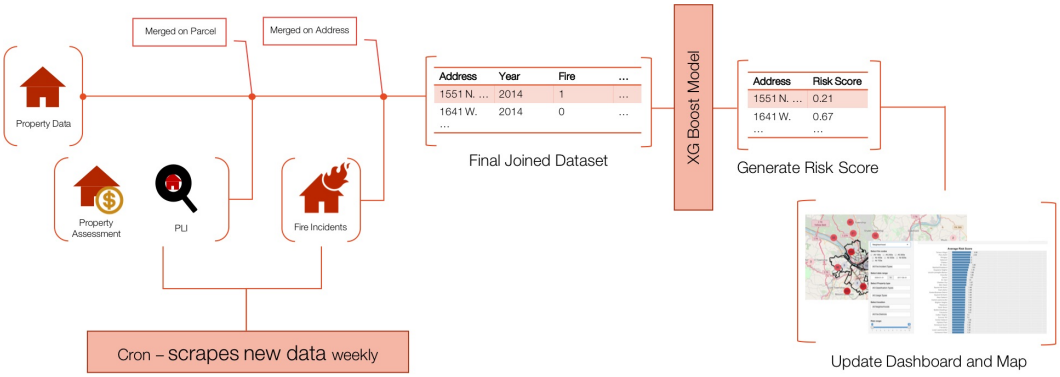

Figure 4: Deployed Commercial Model Pipeline

[14] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L. Addison. 2015. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1909–1918. DOI:http://dx.doi.org/10.1145/2783258.2788620

[15] Michael Madaio, Shang-Tse Chen, Oliver L Haimson, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, and Bistra Dilkina. 2016. Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 185–194.

[16] United Nations Department of Economic and Social Affairs. 2018. Poverty Thresholds. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. (Population Growth 2018). Accessed on 10-5-18.

[17] Bhavkaran Singh Walia, Qianyi Hu, Jeffrey Chen, Fangyan Chen, Jessica Lee, Nathan Kuo, Palak Narang, Jason Batts, Geoffrey Arnold, and Michael Madaio. 2018. A Dynamic Pipeline for Spatio-Temporal Fire Risk Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 764–773.

Appendix

4.1 Fire Risk Model Pipeline

See Figure 4 for a visualization of the deployed model for fire risk prediction, based on the Firebird framework [15].

4.2 Performance Metrics

For our 6 performance metrics (accuracy, AUC, kappa, precision, recall, F1) across each model iteration, we show the raw values in Figure 5, the mean values in Figure 6, and cumulative standard deviation in Figure 7.

4.3 Precision and Recall at Top K

In Figure 8, we rank the properties from the highest risk score to the lowest, and then calculate the precision and recall of our model at each of the k highest risk properties.

4.4 Variance in Feature Importance

In Figure 10, we show the features from the Fire Codes with the most variance in feature importance across model iterations, including, the "EMS call, excluding vehicle accident with injury", "smoke

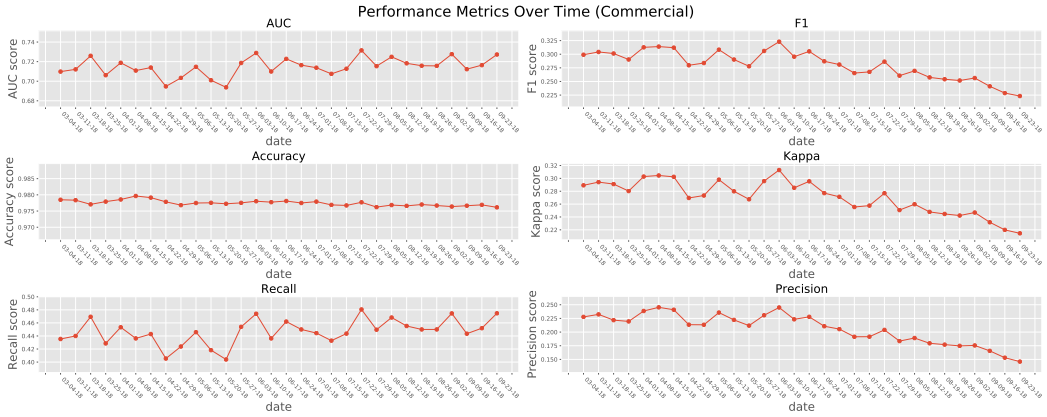

Figure 5: Raw value of the AUC, F1, Accuracy, Kappa, Recall, and Precision

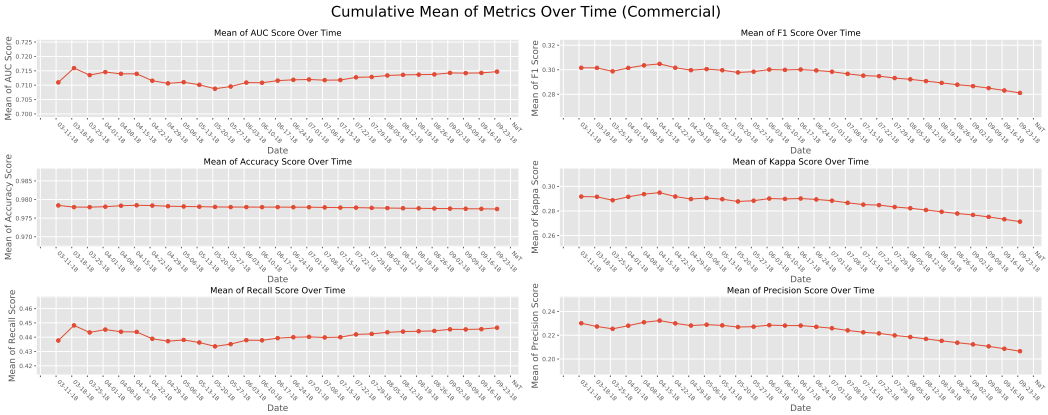

Figure 6: Cumulative Mean of the AUC, F1, Accuracy, Kappa, Recall, and Precision

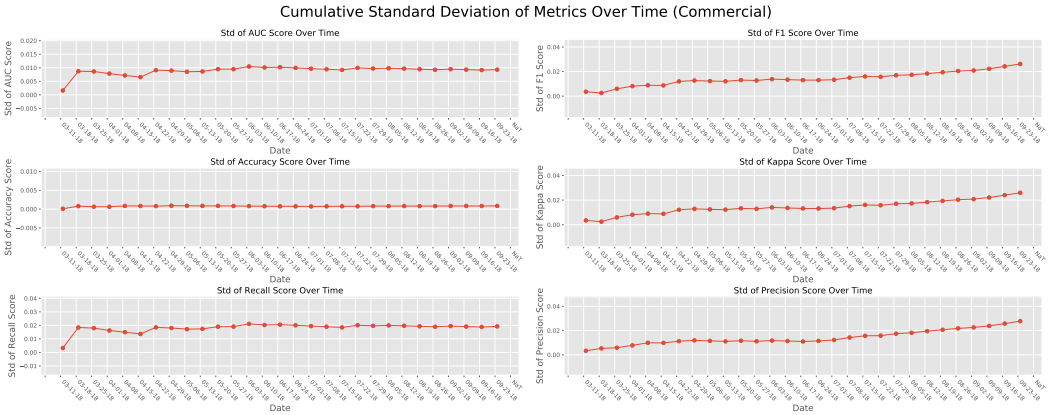

Figure 7: Cumulative Standard Deviation of the AUC, F1, Accuracy, Kappa, Recall, and Precision

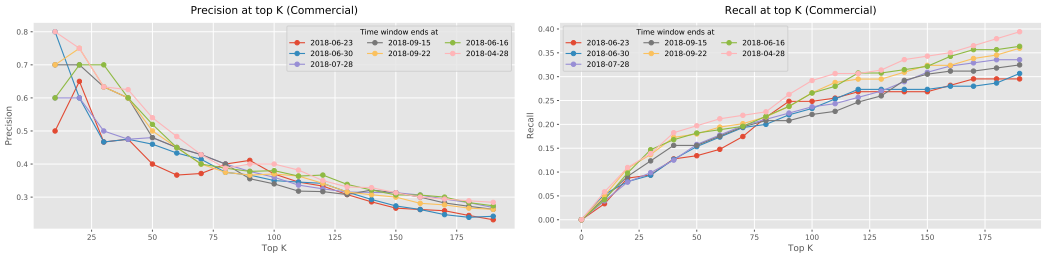

Figure 8: Precision and Recall of the Top-k Properties

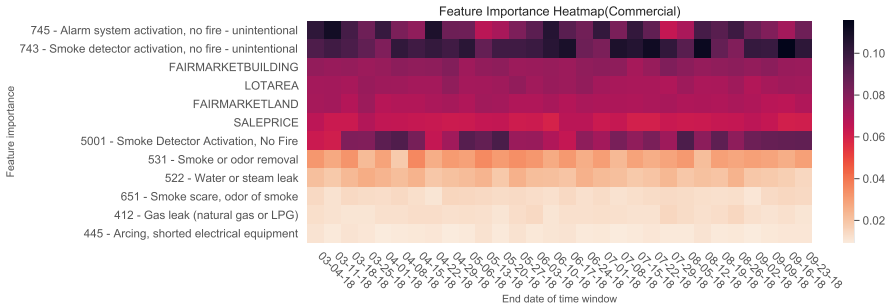

Figure 9: Feature Importance of Top 12 Features Over Time

detector activation no fire - unintentional (743)", "dispatched and cancelled on arrival (6111)", "Alarm system activation with no fire - unintentional (745)", and "medical assist for EMS Crew (311)".

4.5 Risk Group Proportions

Figure 11 shows the proportion of properties in each risk category.

4.6 Simulated Deployment of Residential Fire Risk Model

Much like the commercial risk model, we wanted to understand the robustness and stability of a residential risk model, where the unit of analysis is the census block rather than a single address, due to the fire risk reduction effort – here, fire safety education programs – being done at the community level instead of the property-level. Because this model was not deployed, we simulated the weekly

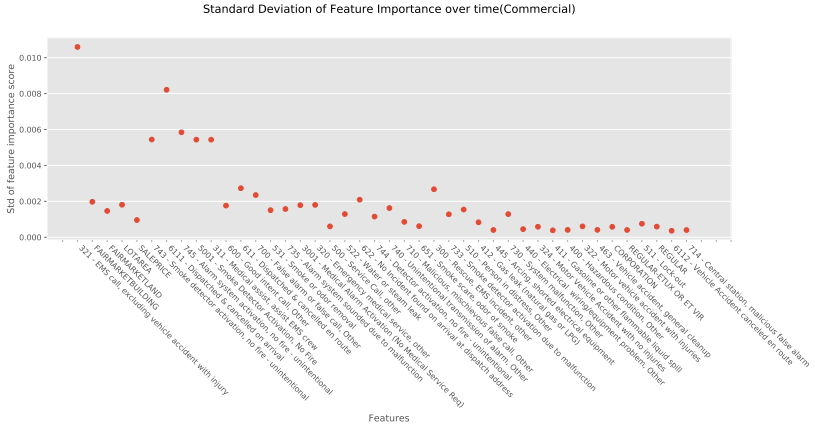

Figure 10: Standard Deviation of Feature Importance over 20 weeks

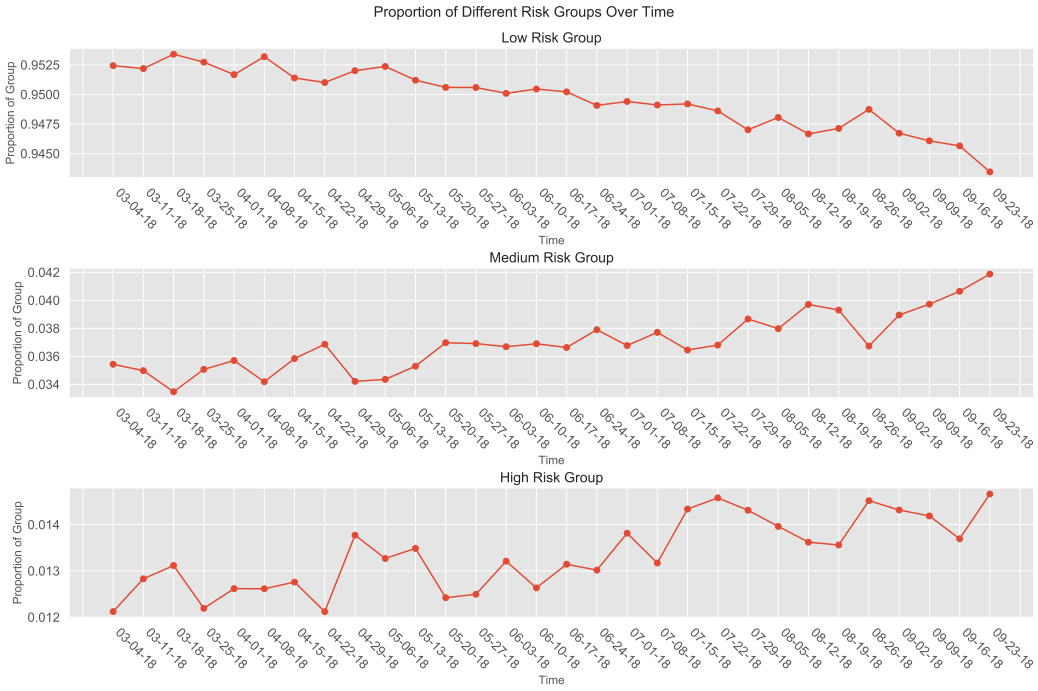

Figure 11: Percent of Properties in Low, Medium, and High Risk Categories

Table 2: Fire Incidents for Simulated Deployment of Residential Model

Risk Score Category	Total Census Blocks	Code 111s	Code 111-118	Code 100-199s	Any Fire Code
> 0.5	283	166 (58%)	195 (68%)	204 (72%)	207 (73%)
< 0.5	59	10 (15%)	14 (23%)	16 (25%)	18 (28%)

iterations of the model by limiting the data at each iteration to only use the data available prior to the date of that simulated weekly iteration, following [1]. In Table 2, we display the results of this simulated residential model, finding that approximately 68% of the high risk census blocks had building fire incidents. However, as nearly 83% of census blocks had a prediction probability greater than 0.5, this suggests a more nuanced approach to risk score discretization may be required. Additionally, because census blocks may contain hundreds of residential properties, there may be multiple fire incidents, suggesting that a binary classification approach may not be the most appropriate or effective. For our 6 performance metrics across each simulated model iteration, we show the raw values (Figure 12), the mean values (Figure 13), and the standard deviation (Figure 14).

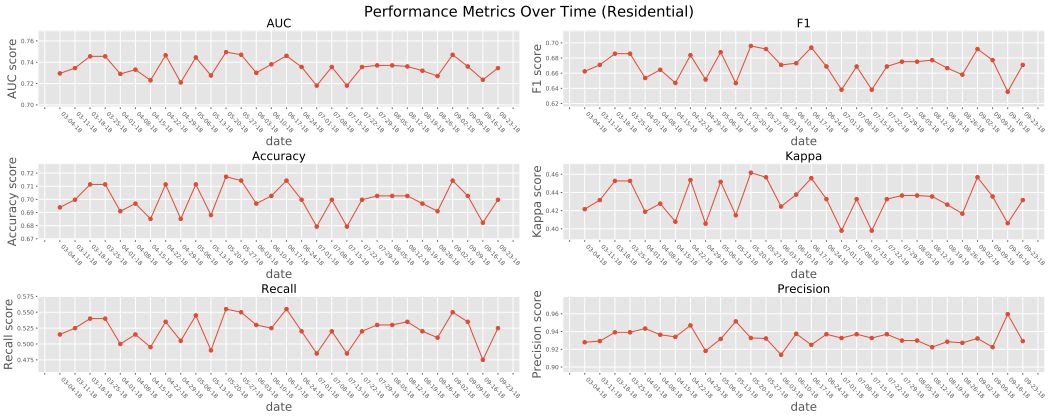

Figure 12: Raw Value of Simulated Residential Deployment Over Time

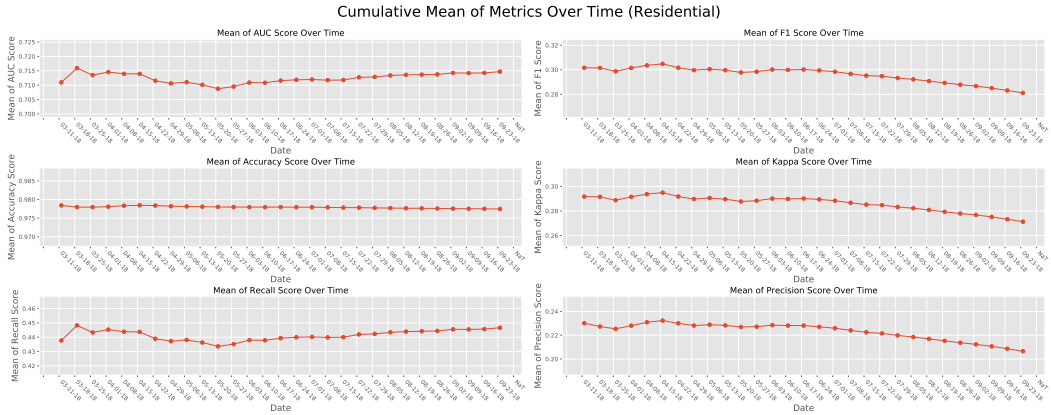

Figure 13: Cumulative Mean of Simulated Residential Deployment Over Time

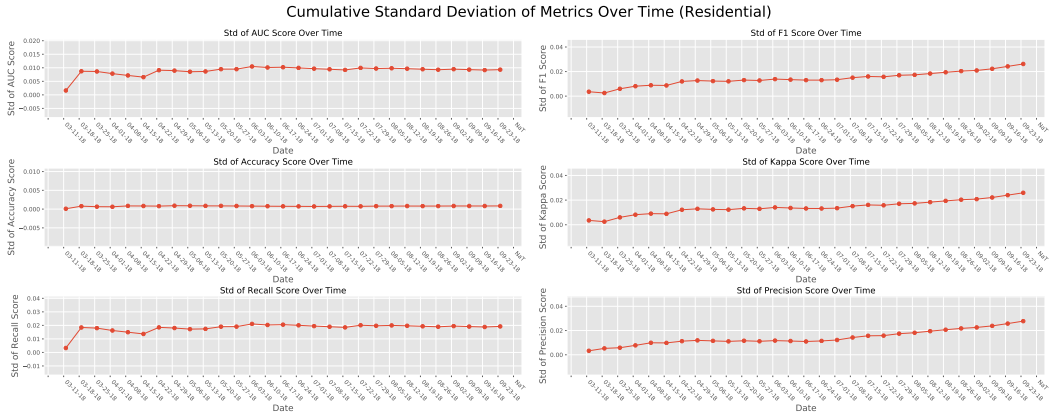

Figure 14: Cumulative Standard Deviation of Simulated Residential Deployment Over Time