FIRE: Fire Insights & Risk Evaluation:

Phase II Multifamily Dwelling Analysis

August 2022 – April 2023

By Margaret Catherman, Contributors: Project Coordinator: Bea Wang; Team Lead: Matt Hinds-Aldrich; Team Members: Margaret Catherman, Alex Dreisbach, Chiyoung Lee, Sanal Shivaprasad, Yvonne Zhang; DCFR Representatives: Tom Burrell, Appraisal and Planning Chief; Dung Nguyen, Management Analyst.

Objective

The objective of Fire Insights & Risk Evaluation, or FIRE II, was to engineer a fire risk prediction model using machine learning and historical data to provide DeKalb County Fire Rescue (DCFR) fire risk rankings and analysis of multifamily property in Dekalb County, GA. The findings will be used by DCFR in their community risk assessment to identify *which* multifamily properties to prioritize for community risk reduction campaigns, as well as *what* fire risk reduction methods are most relevant to these communities and should be included in the outreach, and finally, *when* are the optimal time to conduct these campaigns.

Summary

This was the second phase of the Fire Insights and Risk Evaluation analysis by ATLytiCS volunteers, in collaboration with analysts at DCFR. Phase I, completed in 2022, used predictive modeling to assign fire risk levels to single-family residential properties based on descriptive information about the properties. The objective of Phase II was to design a similar fire risk prediction model and analysis for multifamily properties. As with previous studies, we incorporated descriptive data on the multifamily complexes' buildings, such as area, perimeter, and acres. Additionally, this phase also examined if there was a correlation between metrics reflecting resident's behavior and sentiment in predicting fire risk levels, something rarely included in previous studies. For this, we incorporated two features: Google ratings, to measure residents' sentiment; and to measure residents' behavior, call types such as EMS, Service Calls, and Canceled, good intent, from the fire incident data. Unfortunately, Google ratings was found insignificant in predicting properties' fire risk. The call types, however, were highly correlated with fire occurrence, and instrumental in increasing the model's performance. The model produced a table ranking multifamily properties by their probability of fire: High, Moderate or Low. This table will be incorporated into a dashboard used by the DCFR to identify high risk complexes to target with fire risk reduction educational resources through area schools and community centers. In addition to offering metrics to determine *where* to target outreach, findings of this study may be used to determine *which* fire prevention messages and efforts are most needed in specific communities, as well as *when* community risk reduction (CRR) campaigns should be delivered, based on historic trends in fire incidents.

Background

One of the earliest data driven fire risk analysis of an urban area was conducted in 2013 by the mayor's office in New York City. It replaced focus group assessments by fire fighters with data analytics, and resulted in significant improvement in risk assessment of properties. [3] This and subsequent projects integrated data from numerous local government agencies that had historically been housed in separate siloes. The shift to incorporating data from other agencies for predictive modeling of urban fire risk was a byproduct of the smart city movement, described as, "[the] Smart Cities Initiative takes a forward-looking creative approach to bringing people, policy and technology together to significantly improve the quality of life for metropolitan area citizens." [11]

Since the New York City initiative, other risk based, fire prediction projects have been launched in numerous cities, primarily for commercial properties, but in some instances for residential and multifamily properties. Many projects are

based on methods used in previous analysis, usually with some modifications to improve upon earlier results. Each project uses predictive modeling to rank properties' fire risk level, in most cases using local government data such as parcel, shapefiles, crime rates, building permits, licenses, fire incidents and EMS calls, along with demographic census data. One innovative exception is the use of satellite imagery of properties in the Portland, OR, analysis, which improved the model's performance, while eliminating the time-consuming challenge of joining various data sets. [1] Several projects shared detailed information on their models and important features, as noted in Figure 1.

Fire Risk Study	Model	Important Features
FIRE Phase 1 ATLytiCS/DCFR, DeKalb County, GA (2022)	Random forest	Lot acres, property value. [2]
Fire Underwriters, Canada (2019)	XGBoost	Building area, lot size. [3]
Satellite Imagery, Portland (2018)	Satellite imagery	Satellite imagery. [4]
Pittsburgh, (2017)	XGBoost	EMS calls, lot area, fair market value of building. [5]
Firebird, Atlanta, (2015)	Random forest	Floor size, land area. [6]

Figure 1. Similar Urban Fire Risk Studies

These projects were referenced for guidance and inspiration as we considered new data sources, models and features to incorporate into our analysis, in an effort to improve the model's performance over that of earlier models. We were intrigued by the simplicity and accuracy of satellite imagery used in the Portland project, and researched the feasibility of including this in our analysis. We were also interested in data reflective of the multifamily property residents, as it is people that cause fires. For this we researched the use of EMS and other call types in the fire incident data, Better Business Bureau Ratings, and crowd source ratings in Google Places API. Of these options, we selected all call types and Google Places API to incorporate into our analysis.

The findings of fire risk analysis have different applications, depending on the property type studied. For commercial properties, the predicted fire risk levels are used to identify and prioritize budlings for fire prevention inspections. For residential properties, risk levels are assigned at the parcel, block group or census tract level. Fire departments use the risk level rankings as part of their community risk assessment (CRA) to select *which* high risk areas to target with community risk reduction (CRR) efforts, such as "door knock" campaigns, used by DCFR, in which fire fighters travel house to house educating residents on risk reduction measures and providing smoke detectors where none are present. For multifamily properties, as is the case with this project, DCFR provides CRR outreach at schools near apartment complexes at high risk. CRR campaigns can be created at the local level, or pre-written resources are available to fire departments at the national level, from organizations such as the US Fire Administration, a division of FEMA. [7]

Community risk assessment findings sometimes identify populations in the community with unique fire risk circumstances and safety concerns. In these cases, insights from the analysis may be used to design risk reduction efforts tailored to the needs of these specific segments of the community, or as supporting documentation for grant applications. For example, in 2016, Gwinnett County (Georgia) Fire and Emergency Services conducted a community risk assessment of older adults. The study identified slips, trips and falls, fires and poisoning to be among major risks. Based on these findings, GCFE implemented safety visits to reduce home injury risks for older residents through on-site education. In another instance, the Worcester (Massachusetts) Fire Department's fire risk analysis revealed nearly 24% of cooking fires were coming from four properties that housed low-income, older adults (2015). Using a fire prevention and safety grant, the WFD engaged in a collaborative effort to replace electric coil burners with burners using temperature control technology to prevent auto ignition. 800 electric stoves were retrofitted, and on-site safety education was provided to 900 residents. No stove top fires occurred in any of the units with the new smart technology burners for the six months following the program. [7]

In a similar manner, the findings of FIRE II will be used by DCFR in its community risk assessment. It will help by identifying *which* areas of the community are in greatest need of outreach, and also *what* issues should be included in CRR campaigns to address these communities' most urgent needs. These campaigns are supported by DeKalb County Fire and Rescue's Public Education unit. In addition to the fire risk reduction efforts mentioned earlier, the department's fire safety educators conduct training, lectures and fire safety demonstrations in elementary schools, homeowner associations, businesses and senior centers, free of charge. Topics include the common causes of fire and fire injuries in the home, fire prevention steps, and what to do in case of fire. These programs target the most vulnerable to fire--young children and older adults. Other community outreach resources include Fire Truck Demos, Fire Station Visits, Fire Extinguisher Demos, Basic Fire Safety Demos, Child Seat installations and Smoke Alarm Programs. [18]

This project is a collaboration between DCFR and ATLytiCS. The DCFR is an all-hazards emergency response agency protecting 735,000 residents over 260 square miles with over 700 full-time firefighters. [2] This collaboration is an example the county's movement away from data "silos' within its many agencies and departments, and towards sharing data to

expand the potential for innovative and creative solutions to meet the needs of the community. [17] ATLytiCS provides Atlanta-based nonprofits with analytics and insights to help fund humanitarian initiates within the local community. It enables skilled professionals to give back to their communities, working with data modeling to assist nonprofits and government agency initiatives. [16]

Approach

The approach to this analysis included:

- Domain knowledge: Review of summary papers and code for similar projects; inclusion of DCFR staff in meetings; and consultation with experts.
- The data: Exploratory data analysis and visualization of four data sets, filtered for DeKalb County, multifamily: shapefiles, parcel data, fire incident and crowdsource ratings.
- Joining the data using parcel ID, geometric coordinates and address to produce observations of 1046 multifamily properties.
- Feature engineering and selection, to include metrics reflecting the properties as a whole, resulting in 50 features used in the model for Approach B.
- The predictive model: Selection and fitting of six models, including parameter tuning with a grid to optimize performance.
- Findings: Visualization and presentation of the model's results, including a deliverable: a table ranking DeKalb multifamily properties' fire risk level, to be used by DCFR in assessing communities for fire risk mitigation efforts.
- Code for the project was written in Python, with visualizations in Tableau.

Conclusion and Impact

The DeKalb County Fire & Rescue (DCFR), like most municipal fire departments, conducts community risk reduction outreach for residential and multifamily properties to educate residents on fire risk reduction measures. Prior to FIRE II, DCFR's multifamily community risk assessment relied primarily on intuition. In 2022, ATLytiCS partnered with DCFR to launch FIRE II for multifamily properties, to identify and prioritize properties for community risk reduction educational efforts, using statistical theory, machine learning, county parcel and fire incident data for the period 1/2016-6/2022.

It is said that men, women and children cause fires. For this reason, metrics reflecting residents' behavior were incorporated in the analysis, with the inclusion of all call types from the fire incident data. A strong correlation was found between a complex's fire risk level and measures reflecting the residents' behavior, particularly Call 300s: EMS, Call 500s: Service, Call 600s: Canceled, Good Intent. To prevent "leakage", only one call type per day per complex was used. The second most important group of features after calls for predicting a property's fire risk level were those regarding the complexes' size, as measured by acres, perimeter, and area. In theory, greater square footage may increase the opportunity for fire on its own; however, it is possible that because larger complexes house more residents, fire risk increases with size due to an increase in occupancy count, as it is people that cause fires.

This model performed significantly better than models in other urban fire risk studies, with an AUC of 95%, accuracy of 89%, precision and recall at 89% each, and a kappa of .77, or "Substantial agreement." FIRE II ranked 1046 multifamily properties in DeKalb County, identifying property risk levels as High for 420 properties, Moderate for 167 properties, and Low for 459 properties. A detailed analysis at the district and individual complex level is available in Tableau for DCFR, including an interactive map and dashboard, as seen in Figure 2. FIRE II integrated and visualized fire incidents, property information and risk scores to enhance DCFR's community risk assessments. Together, the rankings and analysis provide DCFR with insights into three components of their community risk assessment: *which* multifamily complexes should be included, *what* fire prevention messages are most relevant, and *when* are the best months to deliver these messages.

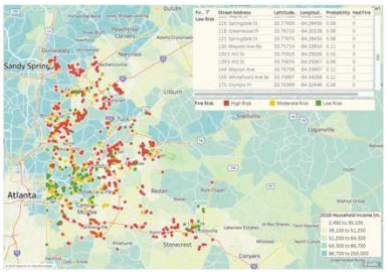


Figure 2. Fire Risk Dashboard with interactive map.

Methods

The data: This project encompassed exploratory data analysis and visualization of four data sets from DeKalb County government, filtered for multifamily properties: fire incident data from Image Trend SaaS RMS (775,910 x 66); parcel data from Dekalb County Property Appraisal (16,234 x 576); shapefile data from Dekalb County Property Appraisal (243,839 x 89) and crowdsourced ratings from Google Places API (1023 x 9). The fire incident data covers 775,910 call records from 1/2016 to 6/2022, along with the callers' address, geolocation coordinates, and detailed information about the call type and reported fire, if one had occurred. Among the observations, 2174 reflect calls reporting a fire. The parcel data represents 1046 multifamily property addresses that encompass 1363 parcel IDs, with observations of 16,234 individual buildings within the complexes, owned by 758 corporations. The data's 578 features include descriptive information about individual buildings within each complex, as well as totals for the complex as a whole, such as area, square footage, year built, type of HVAC system, replacement cost new, number of buildings and floors, building heights, and acres. The shapefiles contain similar information, as well as the geometric coordinates for the boundary of each parcel, needed to join the parcel and fire data. Google Places API was used for its ratings of the properties.

As the goal was to rank fire risk levels for multifamily complexes as a whole, and not the individual buildings in each complex, the data was grouped by each multifamily property's address, resulting in observations of 1046 multifamily properties in DeKalb County, 502 that had one or more fire between 1/2016 and 6/2022. The majority of the properties were "211 Apartment-Garden (3 story & under)" at (72%), with the second largest group "201 Apartments < or = 4 units" at (12%); see Figure 3.

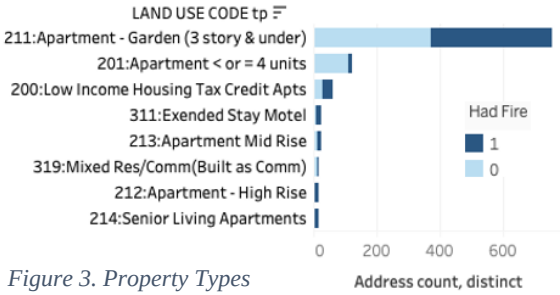


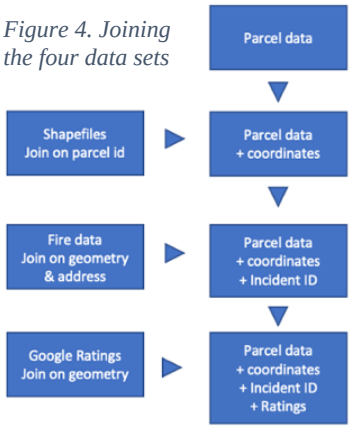
Figure 3. Property Types

The model performed significantly better than models in other urban fire risk studies. One reason for this may be a result of how a fire risk analysis of multifamily properties treats data in comparison to a commercial property analysis. If this were a study of commercial property fires, the 2174 fires would be evaluated out of the 16,234 budlings. Yet because this is an analysis of multifamily properties, the 2174 fires in the 16,234 buildings were grouped within one of the 1046 multifamily complexes. Each complex had between one and 165 buildings. Thus the 2174 fires were assigned to 502 out of 1046 complexes. Each property was then assigned a binary value for whether it had one or more fires over the six-and-a-half-year period of the study: "Had Fire": "0" or "1". No adjustments were made for the overall size or number of buildings per complex, or for the number of fires per complex. This appears consistent with the approach of other fire risk analysis, where there is no evidence of adjustments for the size of buildings or the number of fires per buildings. This approach is adequate for assigning fire risk levels of multifamily properties for fire department purposes, although it might be of interest to consider a linear model for a more granular analysis at some point.

Joining the data: The objective of joining the data was to combine descriptive information of the 1046 multifamily properties from the parcel data, such as the complex's area, perimeter, square footage, number of buildings, and acres, with the geometric coordinates from the shapefiles, with the reported fire Incident ID, information about the fire, date and time from the fire incident data, as well as all call type incident data for Approach B & C; and finally with Google ratings. Once joined, the data was used in the model to predict the probability of a complex having a fire based on descriptive information of the complex, historical data, and resident's actions and sentiment. The first step, joining the parcel data and shapefiles, was easily accomplished on parcel ID. Unfortunately, subsequent joins were not as simple

.

Figure 4. Joining the four data sets



Address proved an inefficient means to join the data, as the parcel data provided only one "official" address and geolocation for property assessment purposes, along with descriptive information on the complex's individual buildings, which range from one to 165, depending on the complex. In contrast, the fire incident data provided one to thirty-five addresses and geolocation per complex. This is because the latter reflects directions for EMS call dispatches, necessitating a specific address and geolocation for each building within the complex. In the case of an apartment complex with only one

building, these addresses and coordinates of the parcel and fire data may match; but as 80% of multifamily properties have multiple buildings, in most cases there is a mismatch. An example: Kensington Apartments. The parcel data provided one address and one geolocation of the property's boundaries, roughly the area seen in Figure 5; and observations for 43 buildings, noted in blue and red. For the same apartment complex, the fire data provided twenty-four addresses and geolocations, noted in red. To complicate matter further, Google Places API provided yet another address and set of geolocation coordinates, which in most cases were not a match with either the fire incident or parcel data.

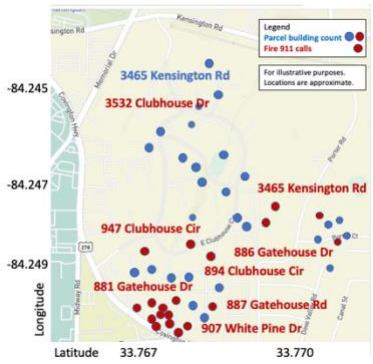


Figure 5. Kensington Apartments

The solution: Join the parcel, fire and ratings data in three steps: first for those few that are an exact match on address, then on geometry using Python's GeoPandas library, first within the boundaries of the parcel, then within a distance of the center point of the parcel. This last step, matching on distance, is "greedy" and indiscriminate. A better alternative is the fuzzy join, which in addition to using distance, included an approximate match of address. Substituting the fuzzy join in place of matching on distance alone increased the model's performance.

Three approaches were used in this analysis, discussed in detail, below. These different approaches impacted the joining of the data. Approach A joined parcel data only with observations of "Call 100s: Fire" in the fire incident data. Approach B & C joined parcel data with all call types. Of the three, Approach B & C allowed for a more accurate and complete join of the data sets, as they provided geolocations and information for over 23,633 distinct fire incident addresses, compared to Approach A, which had information on only 1714 distinct fire incident addresses. All approaches mapped the properties from the fire data to 1072 multifamily complex addresses in the parcel data.

Feature engineering and selection: Once the data was joined, features were grouped by individual multifamily complex. To ensure data represented each complex as a whole, the sum and max were calculated, in addition to the median, depending on the feature. This step is important when working with multifamily properties, as complexes range in size from one to 165 buildings. For example, using only the mean for area will result in the area entered for a complex with one 20,000 square foot building being equal to the area entered for a complex with ten 20,000 square foot building! Not surprisingly, the model performed significantly better when calculations reflecting the complex as a whole were included; in fact, these features were among those most influential for predicting the likelihood of a property having a fire. For Approach A, thirty-one features were selected: Google rating, and thirty attributes representative of various aspects of each complex. The median was calculated for each feature, the sum was calculated for ten features, and the max calculated for one. Combined, forty-two features were used in the model. Features were limited to those with less than 10% missing values, with the exception of google Ratings, which had 14% missing. The mean was entered in place of missing values. The data was divided into training (75%) and test (25%) sets for use in the model.

It is widely agreed that fires are caused by men, women and children. We therefore wanted to go beyond descriptive data about the buildings in the complexes, and incorporate data about the people living there. To achieve this, we included features from two data sources that captured information about residents: Google Ratings, noted above, to get a measure of resident's sentiment. To get a measure of what was happening among the residents, or their behavior, we included all call types recorded in the fire incident data. We broke our analysis into three approaches. All used the descriptive information about each complex from the parcel data, mentioned above, Google Ratings, and observations regarding the dependent variable, "Call 100s: Fire," which was converted to a binomial and renamed "Had fire "(0,1) The approaches differed, however, in that Approach B & C included all call types, with similar feature engineering for each as that applied to "Call 100s", as follows: Each call type was rounded to hundreds to make a series. For example, Call Types 311, 321, 331 and 322 were rounded to 300. Next, each call series was converted to a binomial (0,1). Observations were grouped by the multifamily complexes address from the parcel data, and a count for each call series per complex was tallied. In this manner, these seven call types were added to Approach B: Call 200s: Overpressure rupture, Explosion, Overheat- No Fire, Call 300s: Rescue & EMS, Call 400s: Hazardous Condition- No Fire, Call 500s: Service Call, Call 600s: Canceled, Good Intent, Call 700s: False Alarm False Call, 900: Special Incident Type, Citizen complaint. This brought the total features used in the model for Approach B to 50. Approach C started with the features used in Approach B and omitted independent variables highly correlated with each other, discussed in detail, below. This resulted in 25 independent variables used for Approach B. See Appendix D for a complete list of features used in the three approaches.

Steps were taken to prevent feature leakage of data between the various call types and the dependent variable, "Had Fire" (0,1), which was "Call 100s: Fire", converted to a binomial and renamed, as mentioned previously. Only one call

type per complex per day was used from the data, with priority given to the lowest call category, “Call 100s; Fire”. For example: If a call was made on April 3 from The Avondale Apartments reporting a fire (100 series), no other calls from The Avondale Apartments made that day, such as to EMS (300 series), were recorded. The objective was to see if there was a correlation between various call types and “Had Fire” (0,1), independent of calls made related to a fire incident.

Approach B & C provided a more comprehensive analysis of residents’ behavior at each complex, with 107,548 observations, whereas Approach A offered only 2,734. This contributed to Approach B performing significantly better at each stage of the model development process, from joining the data to a more accurate prediction, see Figure 6a.

Regarding features selected that measured residents’ sentiments and behaviors: the various call types from the fire incident data proved significant in predicting a property's fire risk level, while Google ratings did not. This may be due in part to the manner in which the data was collected for each metric. The fire incident data exhibited greater governance, was more scientific and comprehensive, as one would expect, as it provides the link between citizens in need of assistance and care. The fire incident observations were professionally recorded, without ambiguity: a call type was placed due to a particular series of events, such as to report a fire, or a need for EMS, assistance, false alarm, or hazardous conditions. Missing values for critical information, such as date and district, were nonexistent.

In contrast, Google Ratings lacked both governance and a scientific approach, was subjective, and exhibited signs of manipulation by property owners or others, perhaps as part of a marketing strategy. For example, in Figure 5a, we see a high occurrence of variations of the name "omari simba," who flooded this risk level with his comments. Another problem with Google Ratings is resident’s may or may not share their thoughts. The ratings have limitations to access, as it requires the internet, which may exclude some older or economically disadvantaged residents. The result: mixed messaging at each risk level. For example, in Figure 5a, we see mixed messaging for low-risk properties, with positive comments of "wonderful experience," along with negative comments: "wifi was down," & " they doesn care." Average numeric ratings for each risk level were almost identical, which will be discussed in more detail, below.

Figure 5a. Google Ratings low-risk properties word cloud. Notice the high occurrence of variations of "omari simba."

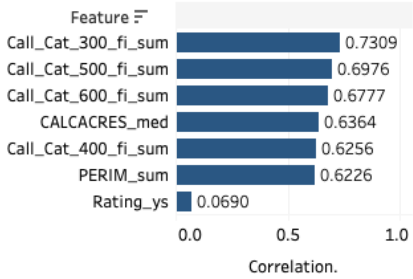


Figure 5b. Spearman's Correlation Matrix rating of top correlated features with the dependent variable “Had Fire” (0,1)

One metric used for determining features to include in the analysis is the Correlation Matrix. It ranks independent variables with strong correlations to the dependent variable, “Had Fire” (0,1) are desirable for inclusion in the model. In Figure 5b, we see Spearman's Correlation Ranking of top independent variables with the dependent variable, "Had Fire “(0,1). Spearman's Correlation is preferred to Pearson's for non-linear variables, as is true for the time series call types used in Approach B & C. Rankings between absolute values of .50 and .75 are considered moderately correlated. Moderately correlated features in the data set included features reflecting residents’ behaviors and actions, as seen in Call 300s, 500s, 600s and 400s, and characteristics related to the complexes’ size, such as acres and perimeter. The crowdsource ratings was found to have no correlation with the dependent variable, receiving a score of only 0.069.

Multicollinearity, or multiple independent variables highly correlated with each other, is a concern with some predictive models, such as linear regression; however, for the classifiers used in this analysis, Random Forest and AdaBoost, multicollinearity is not a concern. Out of curiosity, we ran the model both with and without multicollinear independent variables. Spearman's Correlation Matrix calculates correlation between two variables; but it may miss correlations between multiple variables. Instead, we used variance inflation factor (VIF) values to identify multicollinear variables to omit. VIF values that exceed 5 or 10 indicate a problematic level of collinearity [8]. See Appendix E for a list of independent variables with VIF values less than 10. Only these variables were used in Approach C. Approaches A & B included the highly correlated independent variables with VIF values greater than 10. Interestingly, the model performed

slightly *better* when highly correlated independent variables were retained, as we see in Figure 6A. Therefore, we chose Approach B over Approach C for our primary analysis.

Figure 6a. Performance of the three methods on the test data.

	Calls	Classifier	Accuracy	AUC	Precision	Recall	Kappa
A1	100s	AdaBoost	81%	89%	81%	81%	.63
A2	100s	Random Forest	83%	90%	83%	83%	.66
B1	All	AdaBoost	85%	93%	85%	85%	.70
B2	**All**	**Random Forest**	**89%**	**95%**	**89%**	**89%**	**.77**
C1	All/VIF	AdaBoost	85%	93%	85%	85%	.70
C2	All/VIF	Random Forest	87%	95%	88%	88%	.75

The model: As noted above, the analysis was broken into three approaches. Each included descriptive information about the multifamily complexes from the parcel data and Google Ratings, but differed in that Approach A used only call type 100s, Approach B used all call types, and Approach C was identical to B, except it excluded multicorrelated variables with VIF values over 10. Two ensemble methods for classification were used for each approach, with the results summarized in Figure 6a. The Random Forest classification for Approach B performed the best; therefore, these results will be described in detail, below.

Two ensemble methods for classification were used for this analysis: Random Forest and AdaBoost (Adaptive Boosting). In ensemble learning, a prediction model is built by combining strengths from simpler base models. There are two main steps for this process: developing a set of base learners from the training data, and combining these to form the predictor. [10] Random Forest is an ensemble of decision trees; the algorithm predicts new data by aggregating the predictions of the trees. Each split on each tree uses a random subset of the data, thus decorrelating the trees. [8] AdaBoost differs in that the original data is used, without random sampling; the decision trees have a depth of 1 (i.e., 2 leaves), and predictions made by each tree are weighted. Trees are grown successively, using a slow learning approach. Each new tree is fit to the signal that remains from the earlier trees, and shrunken down before it is used. [8]

A grid was used for both methods to iterate the parameters and recommend the best fit. The recommended parameters for the Random Forest model were: bootstrap: false, maximum depth: 4, maximum features: 9, minimum samples leaf: 3, minimum samples split: 8, n estimator: 46. For AdaBoost, the recommended parameters were: learning rate: 0.6, n estimators: 10, algorithm: SAMME.R

The models' performance: Approach B2: Random Forest Classifier, all call types, performed the best of the three approaches, with an AUC of 95%, accuracy, precision and recall of 89%, and kappa of .77, as seen in Figure 6a. For this reason, Approach B's Random Forest classifier was used for the fire precondition table. A detailed analysis of this model's results follows.

The rankings in Figure 6a were calculated from the ROC curve and confusion matrix. The ROC curve (Receiver Operating Characteristics, a historic name from communication theory) simultaneously displays two error types: true positive and false positive, for all thresholds. True positive (TP/P), also known as recall or sensitivity, is the fraction of properties that "Had Fire" or "1.0", that are predicted correctly. The false positive rate, or Type 1 error, is the fraction of properties with "No Fire" or "0" that were classified incorrectly. In Figure 6b, we see the RUC curve for the Random Forest classifier on the test data. The AUC, or area under the curve, provides the overall performance for the classifier, summarized over all possible thresholds. [8] The AUC for this model is 95%.

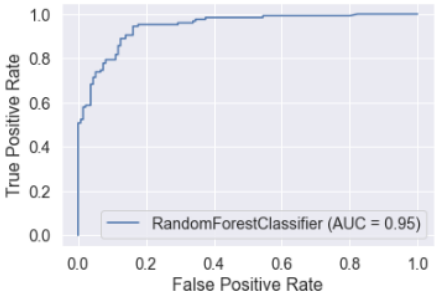


Figure 6b. Random Forest's ROC. AUC = 95%

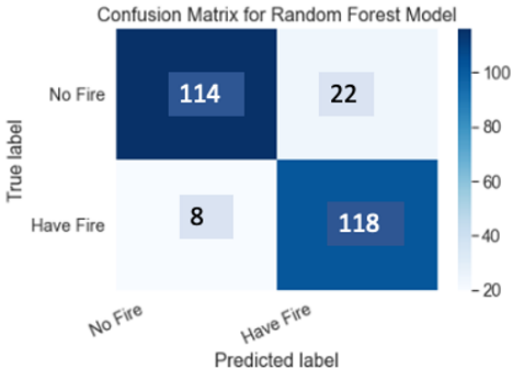


Figure 6c. Approach B Random Forest's confusion matrix

In Figure 6c, the confusion matrix compares the random forest classifier's prediction to the true fire status of the 262 properties in the test data. Precision, recall and the f1-score are calculated from the confusion matrix. [9] Precision *(TP/P**

or TN/N),* also called positive predictive value, is the fraction of *predictions* that are correct. [9] The model's precision rate is higher for properties that did not have a fire, "0.0" or "No Fire", at 94%, compared to those that did have a fire, "1.0" or "Had Fire", at 84%, as noted in Figure 7. These calculations are derived from the confusion matrix as follows: The model predicted 122 (114 + 8) properties would not have a fire, but only 114 properties did not, for 93% precision. The model predicted 140 (22+ 118) properties would have a fire, but only 118 properties did, for 84% precision. Figure 7 provides a table summarizing the random forest classifier's precision, recall, f1-score, and support, or number of observations in the test data.

The model's recall (TP/P or TN/N), also known as sensitivity or true positive, is the fraction of *observations* that are predicted correctly. [9] We see in Figure 7 that the recall was slightly higher for properties that did have a fire (94%), vs. those that did not (84%). Similar to precision, recall is calculated from the confusion matrix, as follows: Among the 136 (114 + 22) instances of a property not having a fire, the model identified 114 correctly, or 84% recall. Among the 126 (8 + 118) properties having a fire, the model identified 118 correctly, or 94% recall.

	precision	recall	fi-score	support
0.0	0.93	0.84	0.88	136
1.0	0.84	0.94	0.89	126
accuracy			0.89	262
macro avg	0.89	0.89	0.89	262
weighted avg	0.89	0.89	0.89	262

Figure 7. Approach B Random Forest's classification report

The f1 score is the harmonic mean of precision and recall, and reflects the number of correct predictions made over all data. [9] The model's f1 score, or accuracy, for predicting properties not having a fire is 88%; for predicting properties having a fire it is 89%, for a combined overall accuracy of 89% among the 262 observations of the test data. The accuracy of the model relates to the accuracy of the fire risk prediction table; thus, the rankings of properties' risk level will have an accuracy rate of ~89%.

The model's kappa was .77, or "Substantial agreement." Cohen's kappa compares the observed accuracy with the excepted accuracy, or random chance, and is calculated as follows:

Kappa = (observed accuracy - expected accuracy)/ (1 - expected accuracy)

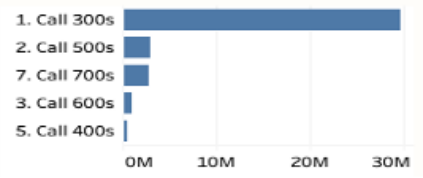
The observed accuracy and expected accuracy are from the confusion matrix; see reference for further explanation of the calculation. [13] Interpretation of kappa is broken into six agreement levels, with scores at the low end of less than 0 termed "Less than chance agreement", and scores at the highest end, between .81 and .99, as "Almost Perfect. Scores between .61 and .80 are ranked "Substantial agreement." [14]

In Figure 7a, we see Approach B2: Random Forest Classifier's ranking of independent variables' importance in predicting the dependent variable, "Had Fire" (0,1). This will vary each time the model is run. Notice how this is influenced by both correlation (Figure. 5b) as well as value counts, Figure. 7b. Example: Call 700 is ranked more influential than Call 400, even though Call 400 has a higher correlation ranking. This is because Call 400 has a much lower value count. Google ratings were not found to be an important feature by the model.

Figure 7a. Random Forest's ranking of important features.

Call 300s	0.2175
Call 600s	0.1247
Call 500s	0.1176
Call 700s	0.1083
CALCACRES	0.1069
AREASUM	0.0710
LINEVAL	0.0640
PRICE	0.0181
CUBICFT	0.0094

Figure 7b. Call totals 1/2016-6/2022



The model in this analysis performed significantly better than those in other analysis. This may be the result of several strategies, such as: incorporating metrics that reflect residents' behavior by including all call types, grouping of individual buildings fires by multifamily properties, including observations that captured the total size of a complex, careful attention to matching data about complexes from different data sources, and using a multi-step join process, that involved joining data by address, geocoordinates within the complex's boundaries, and distance. These strategies are discussed in greater detail elsewhere in this analysis.

Findings & Recommendations

Findings of this study will provide resources for three important components of DCFR's community risk assessment. First, in the identification of *which* multifamily properties have the greatest need of fire risk reduction efforts, based on fire risk rankings from the Fire Risk Prediction Table. Second, the selection of *what* fire risk reduction messages and effort are most relevant to meeting the needs of a specific community, based on analysis of fire incidents, injuries, deaths, fire detector usage and ignition origins. Third, the determination of *when* community risk reduction campaigns should occur, based on the identification of high-incidence months. The following analysis explores these findings by property type or risk level. To aid DCFR in applying these findings to individual complexes, this analysis is available at both the district and individual multifamily property level including address, geolocation coordinates, map location and district in the interactive dashboard accompanying this study.

1. Identification of *which* multifamily properties to prioritize for community risk reduction campaigns, based on predicted fire risk.

Fire Risk Prediction Table: The Fire Risk Prediction table rates 1043 multifamily properties as: High Risk: 420, Moderate Risk: 167, and Low Risk: 459, and includes each complex's address, latitude and longitude and today's date. The table was calculated by applying a probability equation to the model results. Figure 2 illustrates the dashboard that was created by combing the Fire Risk Prediction Table with a map that links the table with the properties' locations.

The importance of targeting properties at high risk is seen in Figure 8, which reflects data for the period 1/2016-6/2022. Here we see high-risk properties accounted for over 90% of fires, building spread and fire related civilian injuries, eight out of the ten fire service injuries, and two out of the four civilian deaths. Of the 20,837 buildings affected by fire spread, 19,481 were in high-risk properties.

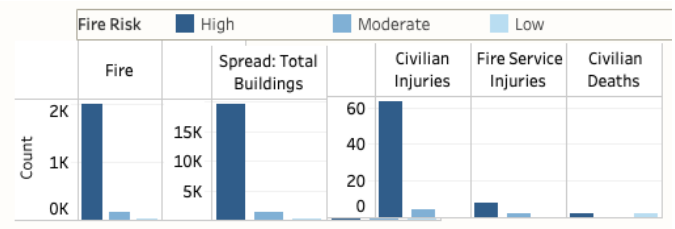


Figure 8. Risk level, fires, spread, injuries & deaths. (1/2016-6/2022)

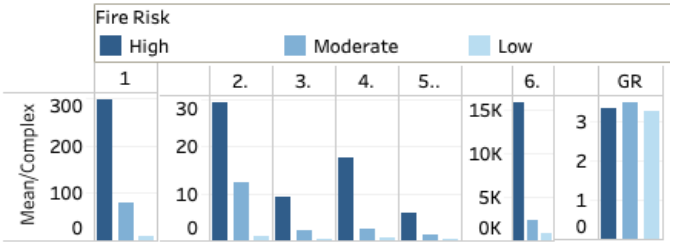


Figure 9. Characteristics of properties by risk level. Numbers correspond to correlation ranking of features to dependent variable "Had Fire" (0,1). 1. Call 300s, 2. Call 500s, 3. Call 600s, 4. Acres, 5. Call 400s, 6. Perimeter, GR: Google Ratings. (1/2016-6/2023.)

Characteristics of High-Risk Complexes: In Figure 9, we see the average measure per complex of the top correlated features, grouped by fire risk rating level. Notice the pattern among these features of descending bars: The higher a property's fire risk rating, the higher the call type count and larger the metrics regarding the property's size. Among high-risk properties, we see the following averages of call count per complex for the period 1/2016-6/2023: Call 300s: 298, Call 500s: 29, Call 600s: 9, and Call 400s: 6. Regarding size, high-risk properties average 17.6 acres, with their perimeters averaging 15,792 linear feet. These values are reduced by about one third for properties with moderate fire risk levels. Values for low fire risk properties are a mere fraction of those of their high-risk counterparts. The lack of correlation between Google Ratings and the dependent variable is illustrated here, as we see the average rating is almost identical for all risk levels at 3.33, 3.52 and 3.26.

2. Identification of *what* issues to consider when tailoring community fire risk reduction campaigns to the needs of a specific community.

After selecting a high fire risk complex to target, consideration should be given to selecting fire risk reduction strategies that addresses the needs of that community. This section offers insight into this endeavor, as it contains observations on the relationships between fire instances, property types, injuries, deaths, fire detector usage and ignition origins. The findings in this section may also help to further prioritize which properties to target. For example, within high-risk properties, priority might be given to complexes with a high repeat fire or injury count.

Multifamily property types with a high percentage of fires *within* their LUC: Among the 1034 multifamily properties in the study, the top three most prevalent Land Use Code types are:

211 Apartments-Garden (3 story and under)	72%
201 Apartments < or = 4 units	11%
200 Low Income Housing Tax Credit Apartments	8%

While the other LUC categories made up less than 3% of all property types, respectively, the higher percentage of repeat fires *within* several of these categories suggest DCFR should consider risk reduction strategies and outreach targeted to these property types. For example, in Figure 10, we see the following categories reported ahigh percentage of fires *within* their LUC category (including repeat fires at the same location):

212:	Apartment- High Rise	96%
200:	Low Income Housing Tax Credit Apartments	87%
214:	Senior Living Apartments	84%

The LUC with the lowest percentage of reported fires in their category was “201: Apartment < or = 4 units”: 10%.

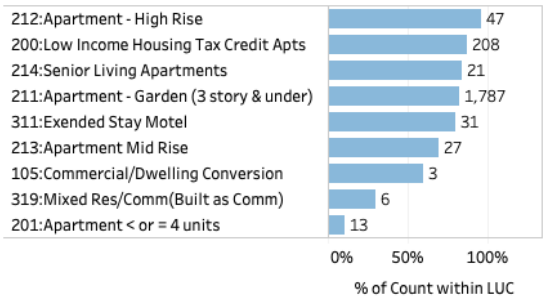


Figure 10. Fires within a LUC (%). Numbers next to bars are fire count, and reflect multiple fires at the same complex.

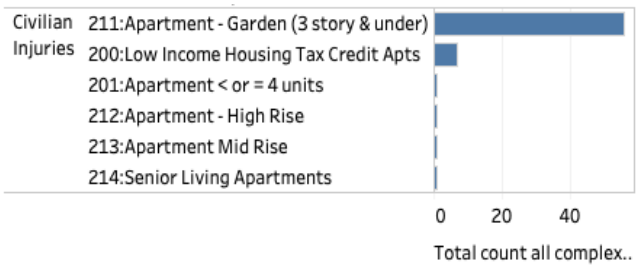


Figure 11. Civilian Injuries by property type

Multifamily property types, **injuries and deaths.** Additional considerations for risk reduction strategies for certain property types should be made in light of injuries and deaths. In Figure 11, we see that among the 67 Civilian Injuries between 1/2016-6/2022, 84% were in “211: Apartment-Garden (3 story & under)”, 10% in “200: Low Income Housing Tax Credit Apts”, with one each in the remaining four property types. Among the ten Fire Service Injuries, 9 occurred in “211: Apartment-Garden (3 story & under)”, and one in “212: Apartment-High Rise”. Of the four Civilian Deaths, two occurred in “201: Apartment < or = 4 units”, and one each in “211: Apartment-Garden (3 story & under)” and “200: Low Income Housing Tax Credit Apts”.

Ignition origin. Statistics for ignition origins should be included in the community risk reduction analysis. In this study, 32% of ignition origins were from within individual apartment units, with the largest single source being “Cooking area, kitchen” at 24%. “Bedroom” accounted for 5%, and “Den, closet”: 4%. Areas shared200: Overpressure rupture, Explosion, Overheat-No Fire, with other residents accounted for 14% of ignition origins, with exterior shared balconies and laundry areas accounting for the largest subgroups. “Building, concealed & exterior wall, engine area+” account for 9% of ignition origins, with concealed & exterior walls and engine areas accounting for the largest subgroup.

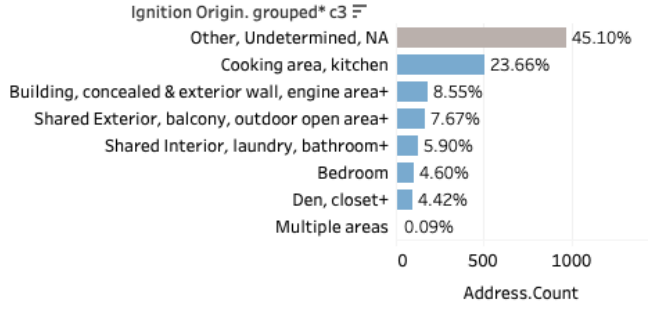


Figure 12. Cooking areas account for 24% of ignitions origins.

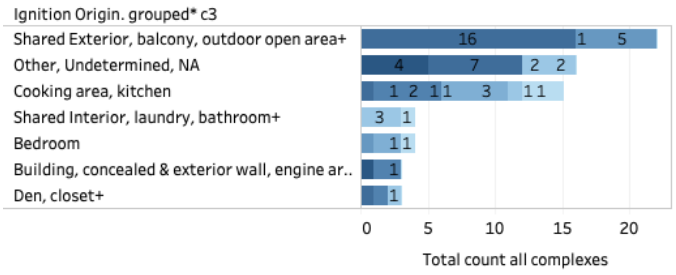


Figure 13. Civilian Injuries by ignition origin. Shading separates incidents; numbers reflect injury count per incident.

Ignition origins, injuries and deaths. While kitchen fires represented the largest single source for ignition origins, as well as accounted for the largest number of civilian injuries at 12 incidents, consideration for other ignition origins should be included in the community risk assessment as well, particularly when examining the relationship between fire origins, injuries and deaths. In Figure 13, we see that among the 67 Civilian Injuries between 1/2016-6/2022, the ignition origin resulting in the most injuries were "Shared Exterior" (33%), followed by "Other" (24%), and "Cooking area" (22%). The largest subgroup of "Shared Exterior" was "Exterior stairway, ramp, or fire escape". Consideration should be given to including fire prevention steps addressing this ignition source, as although there were only two incidents of fires in this group, the injury count was high for each, with 16 injured on 12/11/20 and 5 injured on 12/2/18. More research needs to be done to clarify what happened in these situations. Did the fire *originate* on an exterior staircase or fire escape, or did these injuries *occur* on the exterior staircase or fire escape, perhaps as people were fleeing a fire from within the building? During the time period of the study, there were ten Fire Service Injuries; of these, the ignition sources for 3 were undetermined, 4 were "Cooking area, kitchen", and one each for "Exterior stairway, ramp, or fire escape"", "Bathroom", and "Den". Among the four Civilian Deaths, two ignition sources were in the "Bedroom", one in the "Den" and one 'Unknown". Fortunately, there were no Fire Service Deaths during this period.

Detectors present in communities. Fire detectors were present for 29% of fire incidents, with 10% indicating no detector present, and the remaining 60% were undetermined, see Figure 14. In Figure 15, we see that within specific Ignition

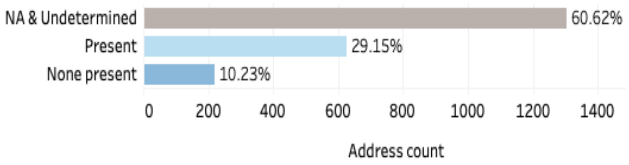


Figure 14. Detectors were not present for 10% of fires.

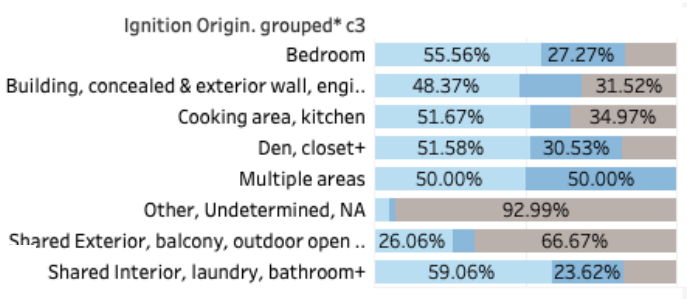


Figure 15. Detectors within an ignition origin, right.

Origins, "Dens, closet +" had the highest percent of detectors not present at 31%, followed by "Bedrooms" at 27%, "Shared Interior, laundry +" at 27%, "Complex Crawl space+"21%, and "Cooking area" 13%. Among Property Types, (not pictured), "201: Apartment < or = 4 units" had 15% of fire incidents without a detector, "214: Senior Living Apartments" had 10%, and "212: Apartments: High Rise" had 6%.

Complexes with high fire occurrence: Fires per complex: 544 complexes reported no fires over the period 1/2016 to 6/2022, 141 reported only one fire, 312 reported 2 to 9 fires, 27 reported 10 to 14, 6 reported 15 or more fires. Within the high-risk ranking, priority should be given to properties with high incidents of fire.

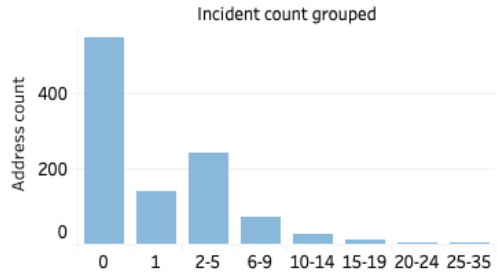


Figure 16. Fires per complex (1/2016-6/2022)

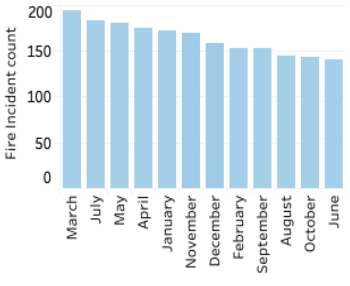


Figure 17. Monthly fire totals, (2016-2022)

3. Identification of *When* outreach efforts should occur

Increase Outreach During High Incident Months: DCFR may want to consider increasing outreach during, or several weeks prior to, the months of March, July and May, as they have the highest historic totals for fires. June, October and August reported the lowest counts. As a general reference, fire incidents were up 13% during the first two years of the pandemic compared with the average for the three years proceeding. Perhaps this was due to more people working from their apartments during this time, or possibly children at home alone when schools were remote.

Next steps

1. The Fire Risk Prediction Table will be incorporated into a RShiny dashboard via an AWS pipeline to enable DCFR to update findings automatically going forward.

Future Study

A. Create a data dictionary and incorporation of features between county government agency data: One challenge of this project was deciding which feature to use to group observations belonging to the same multifamily complex, as was feature selection for the model, due to unfamiliar acronyms used for some features. Data dictionaries would resolve these issues. A dictionary is available for the fire incident data; but not for the Dekalb shapefiles or parcel data. It would be helpful if the county would provide such dictionaries when releasing data. Furthermore, it would help if the county would consider including a few features from the shapefiles in the parcel data, such as the boundary coordinates for each parcel and the census tract, as is done in some other parts of the country. The "doing business as" name for each complex would be helpful as well.

B. Consider incorporating additional county data in the analysis, in particular the county fire inspection report, code complaints and crime records. Other studies, such as those in NYC and Pittsburgh, have included this data in their analysis. An example of what such data looks like for multifamily properties can be seen in the summary for Avondale Village apartments, below. This analysis is from the AJC's "Dangerous Dwelling Series" (2022), and combined data from numerous local sources, to provide the following report:

> **Avondale Village**, 3465 Kensington Rd, Decatur. From 2017-2021, police reported 349 crimes at these addresses, including four homicides, 45 aggravated assaults, 24 robberies and 10 rapes. The complex had fires in February 2017, June 2017, January 2018, May 2019, March 2022 and May 2022. One person died in the May 2022 fire. DeKalb shows 14 code complaints at the Kensington Road address. At the White Pine addresses were 43. At addresses in the 800 block of Gatehouse there were 30. County data show 191 code complaints at addresses on Clubhouse Circle. In 2017 inspections, the county listed 501 code citations at the complex, the most of any in unincorporated DeKalb. A county fire inspection report in August 2022 noted that there was not current annual fire hydrant inspection paperwork; that emergency lighting was not working at the Leasing office, that the fire extinguisher in the kitchen of the leasing office was out of date; and that the property manager could not produce the smoke alarm log. [15]

References

1. Hinds-Aldrich, M. (2020), Pre-Fire: The problems and promise of fire risk prediction. https://medium.com/@dr._matt/pre-fire-the-promise-and-problems-of-fire-risk-prediction-e84986dca131
2. Pawar, S., (2022), DeKalb County Fire Rescue (DCFR) Fire Risk Mitigation Project. https://atlytics.org/dekalb-county-fire-rescue-dcfr-fire-risk-mitigation-project/
3. McGuinness, R., Gilbert, G., Whittamore, Z., Zhong, L., Van Zyl, R., Lee, J. (2019), FUS Building Fire Risk Prediction Validation Project. https://fireunderwriters.ca/assets/img/FUS_Building_Fire_Risk_Validation_Project.pdf
4. Jay, J., (2018), Fire prediction using satellite imagery: exploratory results. RPubs. https://rpubs.com/jonjay/fire_sat
5. Walia, B., Hu, Q., Chen, J., Chen, F., Lee, J., Kuo, N., Narang, P., Batts, J., Arnold, G., Madaio, M. (2017), A Dynamic Pipeline for Spatio-Temporal Fire Risk Prediction. http://michaelmadaio.com/KDD_2018_FireRisk.pdf
6. Madaio, M., Zhang, W., Chen, S., Cheng, X., Haimson, O., Hinds-Aldrich, M., Chau, D., Dilkina, B. (2015), Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. https://poloclub.github.io/polochau/papers/16-kdd-firebird.pdf
7. Community Risk Reduction (CRR) resources for fire departments, US Fire Administration. https://www.usfa.fema.gov/a-z/community-risk-reduction.html
8. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), An *Introduction to Statistical Learning, Second Edition*, *with Applications in R*. Springer Science + Business Media, LLC, New York, NY, pp. 151:153, 342: 344, 352. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
9. Wikipedia, Precision and recall. https://en.wikipedia.org/wiki/Precision_and_recall
10. Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition.* Springer Science + Business Media, LLC, New York, NY, p. 605. https://hastie.su.domains/Papers/ESLII.pdf

11. Minister of Transportation and Infrastructure, "Bill 4 - 2016: Fire Safety Act - Part 6 - Compliance Monitoring." https://www.leg.bc.ca/parliamentary-business/legislation-debates-proceedings/40th- parliament/5th-session/bills/third-reading/gov04-3.
12. Bloomberg, M., Mayor City of New York, Flowers, M., Chief Analytics Officer, Mayor's Office of Data Analytics, (2013), NYC Analytics, NYC by the Numbers Annual Report – 2013. https://a860-gpp.nyc.gov/concern/parent/pc289j734/file_sets/ks65hc82t
13. Cohen's Kappa in plain English. https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

14. Cohen's Kappa, https://stackoverflow.com/questions/11528150/inter-rater-agreement-in-python-cohens-kappa
15. Atlanta Journal and Constitution, Dangerous Dwelling Series, (2022), https://www.ajc.com/news/investigations/dwellings/list-dangerous-apartments/
16. https://atlytics.org
17. https://www.dekalbcountyga.gov/sites/default/files/users/user3566/DeKalb%20County%20Technology%20Strategic%20Plan%20DEC2022.pdf
18. https://www.dekalbcountyga.gov/fire-rescue/public-education

Appendix

A. FIRE Phase 2: Code & data, https://github.com/catherman/FIRE-II/blob/main/ph2_join_w_model.v8_approach_B.ipynb
B. FIRE Phase 2: Visualization: https://public.tableau.com/app/profile/margaret.catherman/viz/FIREIIMultifamilyFireRiskAnalytics/SummaryPublic
C. Parcel data dictionary: OrionCAMADataDictionary. General reference; not exact match with DeKalb parcel data. https://docs.google.com/spreadsheets/d/1kNznf5Pejb1Q5B7fEWYnITmxHAeSSyOn/edit#gid=984329834
D. Fire incident data dictionary: https://docs.google.com/document/d/10gg2d6sv4a7O-Ol63vP1nh4ydww-Ly-91e238_VFjHk/edit
E. List of independent variables used in Approach C, as their VIF values <10: 'CALCACRES_med', 'CONSTR_med', 'Call_Cat_200_fi_sum', 'Call_Cat_300_fi_sum', 'Call_Cat_400_fi_sum', 'Call_Cat_500_fi_sum', 'Call_Cat_600_fi_sum', 'Call_Cat_700_fi_sum', 'Call_Cat_800_fi_sum', 'Call_Cat_900_fi_sum', 'DEPR_med', 'FLRFROM_med', 'FUNCTUTIL_med', 'LLINE_med', 'MSSECT_med', 'NBHD_med', 'OCCUPANCY_med', 'PERIM_med', 'PHYCOND_med', 'Rating_ys', 'STATUS_2_med', 'STORIES_med', 'YR BUILT_med', 'ZIP1_1_med'
F. Complete list of independent variables used in both Approach A & B. Italics for those used in Approach B only: 'ADJRCN_med', 'APRTOT_med', 'AREASUM_med', 'AREASUM_sum', 'AREA_med', 'Address', 'BASERATE_med', 'BUILDING_max', 'BUILDING_med', 'CALCACRES_med', 'CONSTR_med', 'CUBICFT_med', 'CUBICFT_sum',*'Call_Cat_200_fi_sum', 'Call_Cat_300_fi_sum','Call_Cat_400_fi_sum','Call_Cat_500_fi_sum', 'Call_Cat_600_fi_sum', 'Call_Cat_700_fi_sum','Call_Cat_800_fi_sum', 'Call_Cat_900_fi_sum'*, 'DEPR_med', 'DEPR_sum', 'FEATVAL_med', 'FEATVAL_sum', 'FLRFROM_med','FUNCTUTIL_med','INCUSE_med', 'LINEVAL_med', 'LINEVAL_sum', 'LLINE_med', 'LUC_med',LUC_tp', 'MSRANK_med', 'MSRANK_sum', 'MSSECT_med', 'MSSECT_sum','NBHD_med', 'OCCUPANCY_med', 'OFCARD_med', 'PERIM_med', 'PERIM_sum', 'PHYCOND_med', 'PRICE_med', 'PRICE_sum', 'RATE_med', 'Rating_ys'. 'STATUS_2_med', 'STORIES_med', 'YR BUILT_med', 'ZIP1_1_med'