Fire Risk Ratings & Probability Analysis For Residential & Multifamily Properties DeKalb County

By Margaret Catherman

Summer 2024

Deep-Dive GEOID Terminology Used in this Analysis:

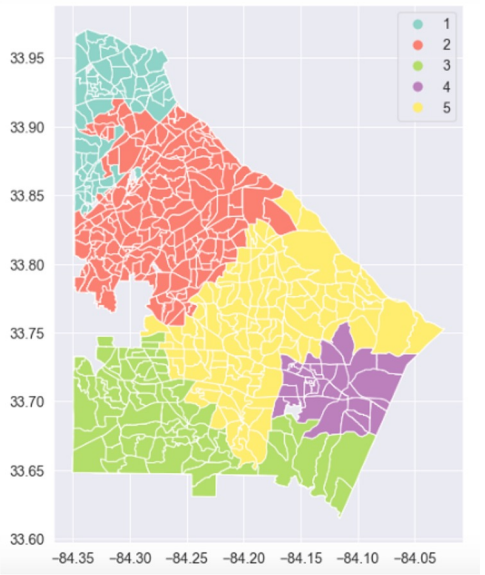
In general, GEOIDs may have 12- or 15-digits total, depending on how detailed the block data is. For example:

	State	county	tract	block
a. GEOID 15 digits, or BG_ID_9	13	089	021823	1927
b. GEOID 12 digits, or BG_ID_6	13	089	021823	1

Throughout this analysis, at times data is referenced at the GEOID 15-digit level, called 'BG ID 9', with the last nine digits of the complete GEOID 15 used as the identifier. For example, "a", above, is noted as BG_ID_9: "218231927".

At other times in this analysis, data is referenced at the GEOID 12-digit level, called 'BG ID 6', with the last six digits of the complete GEOID 12 used as the identifier. For example, "b", above, is noted as BG_ID_6: "218231".

This approach was adopted for brevity as welll as consistency with analysis done in earlier phases of this project. Furthermore, since this analysis is exclusively of DeKalb County, GA, referencing the first six digits of the GEOIDs (state and county), plus the "0" first digit placeholder in tract, is unnecessary & repetitive



GEOID Numbers:

DeKalb County total count:	GEOID 15 digits, or BG_ID_9:	6919
	GEOID 12 digits, or BG_ID_6:	535
Of these, only GEOIDs within DCFR's boundaries are referenced. So, excluding the city of Atlanta, DCFR encompasses:	GEOID 15 digits, or BG_ID_9:	4923
	GEOID 12 digits, or BG_ID_6:	487
Of these:		
Residential homes (1 & 2 family):	GEOID 15 digits, or BG_ID_9:	3392
	GEOID 12 digits, or BG_ID_6:	454
Multifamily properties are:	Address: :	647
	GEOID 12 digits, or BG_ID_6:	225

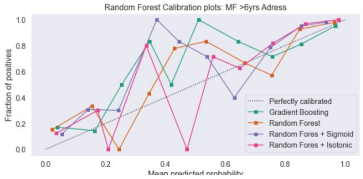
DeKalb Fire Risk Ratings & Probability Analysis
For Multifamily 6 + years of data

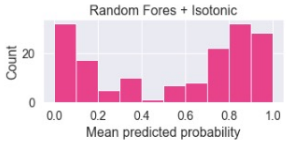
Question: Should the model be fit at the BGID 6 or Address Level?	
Address level **Count: 647**	**Vs. BGID 6 level** **Count: 225**
Preferred approach to fit model by analysts because:	DCFR preferred, as the size of the 225 block groups are more workable for Community Risk Reduction (CRR) campaigns.
Better balance of Fire (0,1): (227, 420) Calculations of probability strong	Imbalanced of Fire (0, 1): (18, 216) Calculations of probability difficult

Recommended Solution:

1. Select features highly correlated with dependent variable. Among these, omit those with correlations greater than 0.85 with each other to prevent collinearity.
2. Fit model & assign risk to the Address level
3. Group these results by BGID 6 and get mean

Results at Address level: Accuracy of 83%, & good precision/recall/f1 for both Fire (0,1)





	precision	recall	f1-score	support
0.0	0.71	0.86	0.78	57
1.0	0.91	0.81	0.86	105
accuracy			0.83	162
macro avg	0.81	0.83	0.82	162
weighted avg	0.84	0.83	0.83	162

	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Classifier MF >6yrs Adr								
Random Fores + Isotonic	0.12	0.38	0.92	0.90	0.82	0.86	0.82	0.62

Conclusion: Making the results available in this manner combines the improved accuracy of the model at the multifamily properties' address level, with the manageability of risk ratings by the 225 BGID 6s, preferred by DCFR. DCFR may focus efforts on high risk BGID 6s. If time permits, individual Multifamily Properties at high risk within moderate or low risk areas may be targeted.

Fire Risk Multifamily Address: *dots*		Fire Risk Multifamily BGID 6: *colors*	
High Risk	360 black	High Risk	168 red
Moderate Risk	63 gray	Moderate Risk	34 orange
Low Risk	220 white	Low Risk	28 yellow

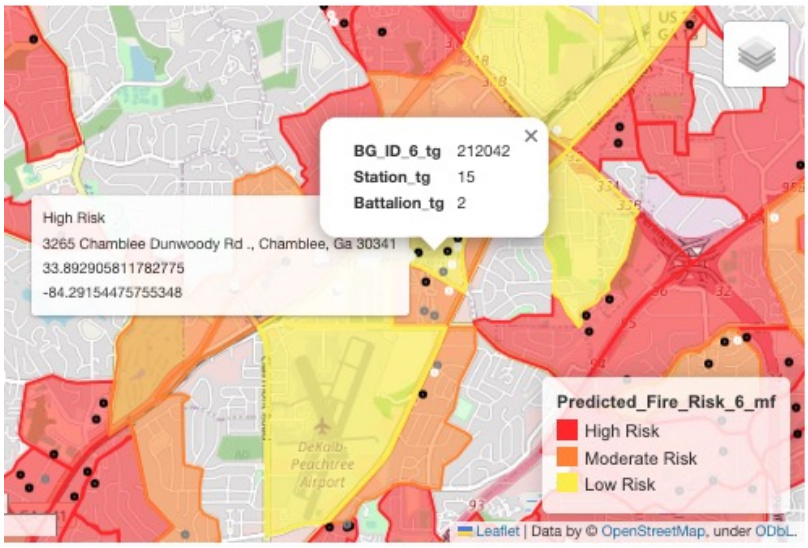


Figure 1: This is a portion of the map illustrating the probability of fire risk with the model fit to a multifamily property address, as described above. Each dot is the center point of one of the 647 multifamily properties. The probability of fire risk at the BGID 6 level was then determined by grouping these results by BGID 6 and calculating the mean. Each of the 225 BGID 6s is colored to reflect this mean, as noted in the legend. Notice 3265 Chamblee Dunwoody Rd. is high risk, although its respective BGID 6: 212042, is yellow, or low risk. If scheduling permits, DCFR may target similar high risk multifamily properties within low or moderate risk BGID 6s.

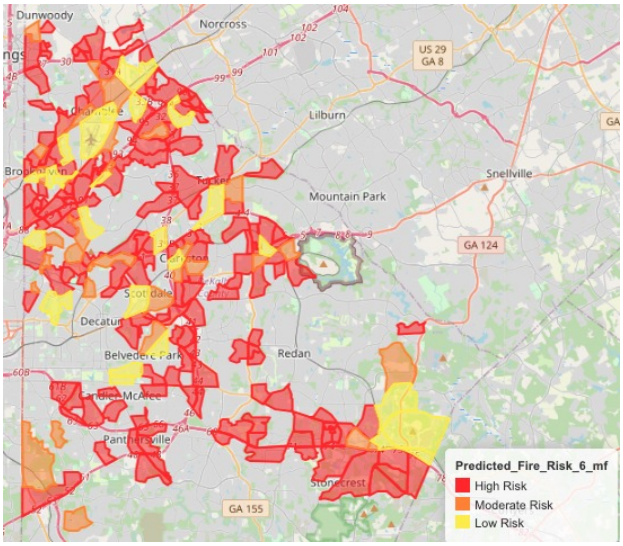
See A1. 6+ yrs. data, SF: BGID_9; Code: 'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb'

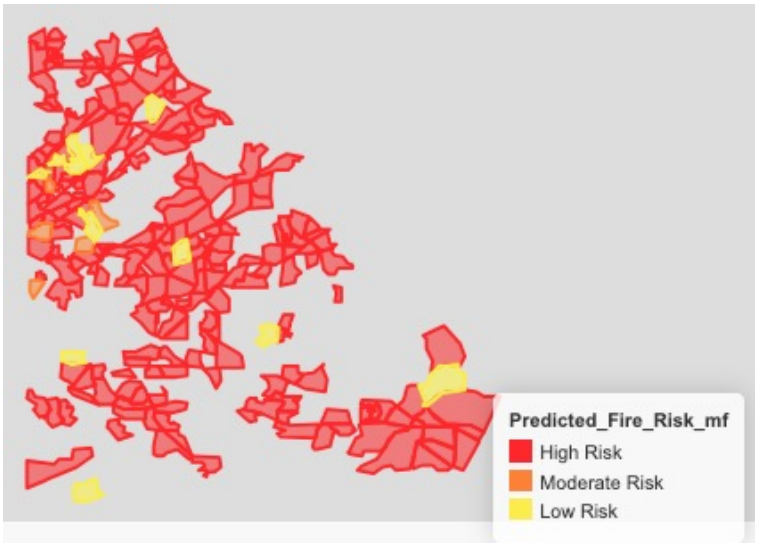
Comparison of Result of Multifamily Approaches

Model fit to Complex Address

Model fit to BGID 6

Models fit to 6 + years of data





>>> *Recommended*: A2. 6+ yrs. data, MF: Address:
Fire (0,1): (227, 420)
Accuracy of 83%, & good precision/recall/f1 for both Fire (0,1)

	Address	BGID_6
High Risk:	**377**	**168**
Moderate Risk:	**64**	**31**
Low Risk:	**206**	**25**

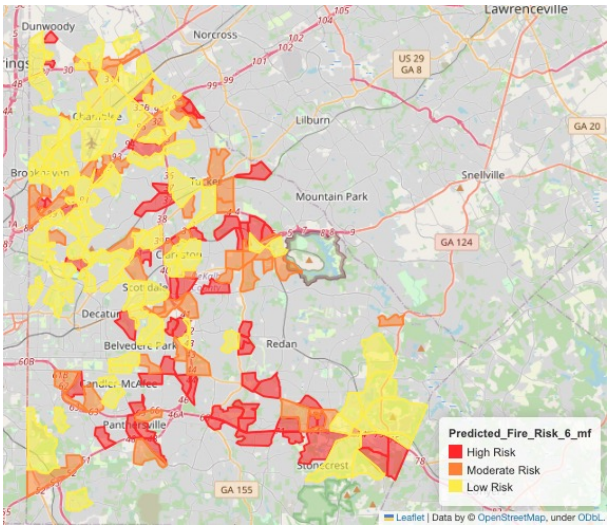
now: ph2_join_w_model.v5_MF-Copy4_LDA_6%2Byrs-Copy5b-Copy26.ipynb
was: ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb

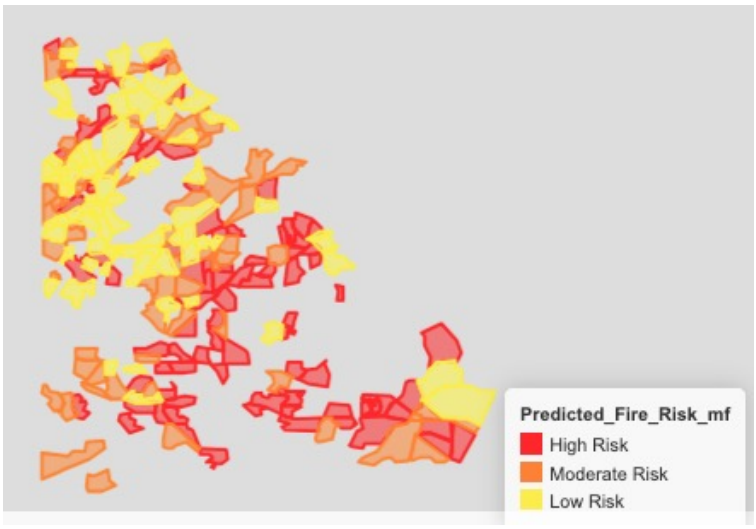
B2. 6+ yrs. data, MF: BGID_6:
Fire (0,1)= (18, 216) Accuracy of 93%, but weak precision/recall/f1 for Fire (0)

	BGID_6
High Risk:	217
Moderate Risk:	12
Low Risk:	5

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy9.ipynb'

Models fit to < 1 year of data





C2. < 1 yr. data, MF: Address:
Fire (0,1)= (467, 180) Accuracy of 75%, & weak precision/recall/f1 for Fire (1)

	Address	BGID_6
High Risk:	98	65
Moderate Risk:	232	98
Low Risk:	317	62

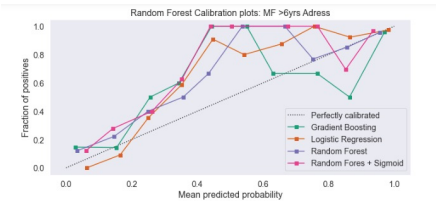
ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy21.ipynb

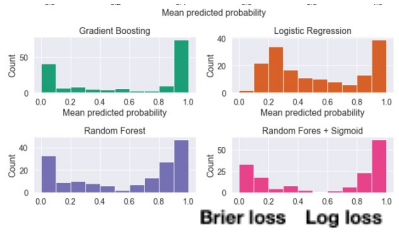
D2. < 1 yr. data, MF: BGID_6:
Fire (0,1)= (108, 126) Accuracy of 61%, & weak precision/recall/f1 for Fire (0,1)

	BGID_6
High Risk:	88
Moderate Risk:	77
Low Risk:	69

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy24.ipynb'

Deep-Dive Comparison of Various Approaches: MF





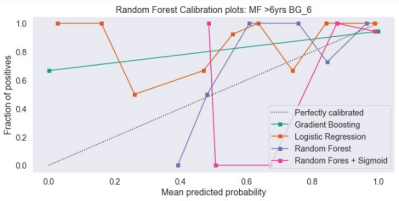

```
12 print(crsf_rfcmf_sigmoid)
```

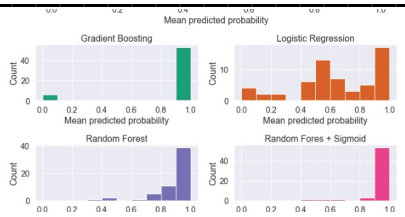
	precision	recall	f1-score	support
0.0	0.74	0.84	0.79	57
1.0	0.91	0.84	0.87	105
accuracy			0.84	162
macro avg	0.82	0.84	0.83	162
weighted avg	0.85	0.84	0.84	162

>>> A2. 6+ yrs. data, MF: Address: Accuracy of 83%, & good precision/recall/f1 for both Fire (0,1)

ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb

Classifier MF >6yrs Adr	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.13	0.43	0.91	0.90	0.84	0.87	0.83	0.64
Logistic Regression	0.15	0.45	0.90	0.93	0.68	0.78	0.76	0.53
Random Forest	0.13	0.42	0.89	0.90	0.84	0.87	0.83	0.64
Random Fores + Sigmoid	0.13	0.40	0.91	0.91	0.82	0.86	0.83	0.64





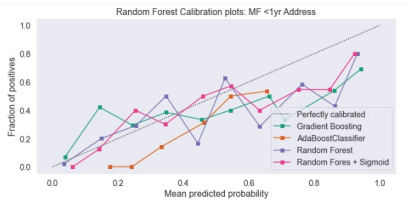

```
4 #rfcmf
```

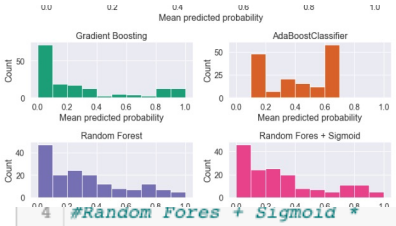
	precision	recall	f1-score	support
0	0.50	0.20	0.29	5
1	0.93	0.98	0.95	54
accuracy			0.92	59
macro avg	0.71	0.59	0.62	59
weighted avg	0.89	0.92	0.90	59

B2. 6+ yrs. data, MF: BGID_6: Accuracy of 93%, but weak precision/recall/f1 for Fire (0)

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy9.ipynb'

Classifier MF >6yrs BG_6	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.12	1.21	0.81	0.94	0.93	0.93	0.88	0.30
Logistic Regression	0.20	0.68	0.77	0.96	0.80	0.87	0.78	0.22
Random Forest	0.06	0.21	0.89	0.93	1.00	0.96	0.93	0.31
Random Fores + Sigmoid	0.05	0.19	0.92	0.93	1.00	0.96	0.93	0.31






```
1 print(crmf) #< 1yr Address
```

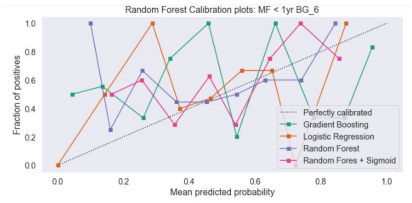
	precision	recall	f1-score	support
0.0	0.80	0.83	0.82	117
1.0	0.51	0.47	0.49	45
accuracy			0.73	162
macro avg	0.66	0.65	0.65	162
weighted avg	0.72	0.73	0.72	162

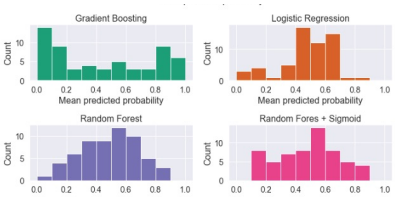
```
4 #Random Fores + Sigmoid *
```

C2. < 1 yr. data, MF: Address: Accuracy of 75%, & weak precision/recall/f1 for Fire (1)

ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy21.ipynb

Classifier MF <1yr Address	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.18	0.54	0.80	0.55	0.47	0.51	0.75	0.34
AdaBoostClassifier	0.17	0.51	0.82	0.53	0.82	0.64	0.75	0.46
Random Forest	0.17	0.48	0.79	0.54	0.47	0.50	0.74	0.33
Random Fores + Sigmoid	0.17	0.48	0.81	0.53	0.44	0.48	0.73	0.31






```
3 #MF < 1yr BG_6
```

	precision	recall	f1-score	support
0	0.52	0.59	0.55	27
1	0.61	0.53	0.57	32
accuracy			0.56	59
macro avg	0.56	0.56	0.56	59
weighted avg	0.57	0.56	0.56	59

D2. < 1 yr. data, MF: BGID_6: Accuracy of 61%, & weak precision/recall/f1 for Fire (0,1)

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy24.ipynb'

Classifier MF < 1yr BG_6	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.37	1.13	0.53	0.50	0.41	0.45	0.46	-0.07
Logistic Regression	0.24	0.67	0.62	0.66	0.59	0.62	0.61	0.22
Random Forest	0.25	0.70	0.62	0.61	0.53	0.57	0.56	0.12
Random Fores + Sigmoid	0.25	0.69	0.64	0.61	0.59	0.60	0.58	0.15

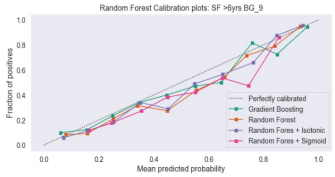
DeKalb Fire Risk Ratings & Probability Analysis
For Residential (1 & 2 family) 6 + years of data

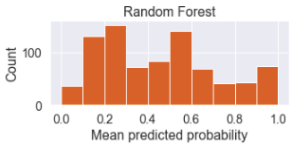
Question: Should the model be fit at the BGID 6 or 9 Level?	
BGID 9 level Count: 3392	**Vs. BGID 6 level Count: 454**
Preferred approach to fit model by analysts because:	DCFR preferred, as the size of the 454 block groups are more workable for Community Risk Reduction (CRR) campaigns.
Better balance of Fire (0,1): (2084, 1308) > Calculations of probability strong	Imbalanced (sparce) of Fire (0, 1): (1, 115) > Calculations of probability difficult

Recommended Solution:

1. Select features highly correlated with dependent variable. Among these, omit those with correlations greater than 0.85 with each other to prevent collinearity.
2. Fit model & assign risk to BGID 9
3. Group these results by BGID 6, & get mean.

Results at BGID 9 level: Accuracy of 75%, & gooc





	precision	recall	f1-score	support
0	0.78	0.81	0.79	521
1	0.68	0.65	0.66	327
accuracy			0.74	848
macro avg	0.73	0.73	0.73	848
weighted avg	0.74	0.74	0.74	848

Classifier SF >6yrs BG_9	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Random Forest	0.18	0.53	0.80	0.68	0.63	0.65	0.74	0.45

Fire Risk Single Family BGID 9: *dots*		Fire Risk Single Family BGID 6: *colors*	
High Risk	945 black	High Risk	59 red
Moderate Risk	1621 gray	Moderate Risk	390 orange
Low Risk	1672 white	Low Risk	112 yellow

Conclusion: Making the results available in this manner combines the improved accuracy of the model at the BGID 9 level, with the manageability of risk ratings by the 400 BGID 6s, preferred by DCFR. DCFR may focus efforts on High risk BGID 6s. If time permits, individual High risk BGID 9s within Moderate or Low risk areas may be targeted.

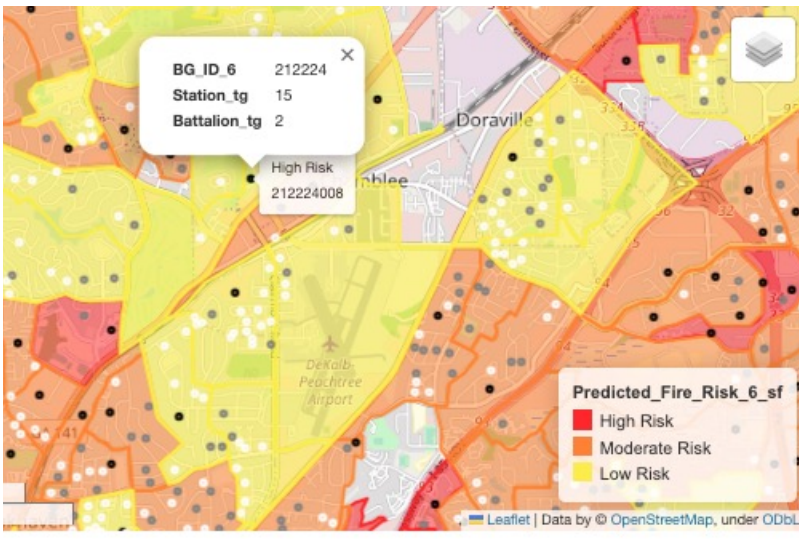


Figure 1: This is a portion of the map illustrating the probability of fire risk with the model fit to BGID 9, as described above. Each dot is the center point of the 3392 BGID 9s. The probability of fire risk at the BGID 6 level was then determined by grouping these results by BGID 6 and calculating the mean. Each of the 454 BGID 6s is colored to reflect this mean, as noted in the legend. Notice BGID 9 "212224002"" is high risk, although its respective BGID 6 is yellow, or low risk. If scheduling permits, DCFR may target similar high risk BGID 9's within low risk BGID 6s.

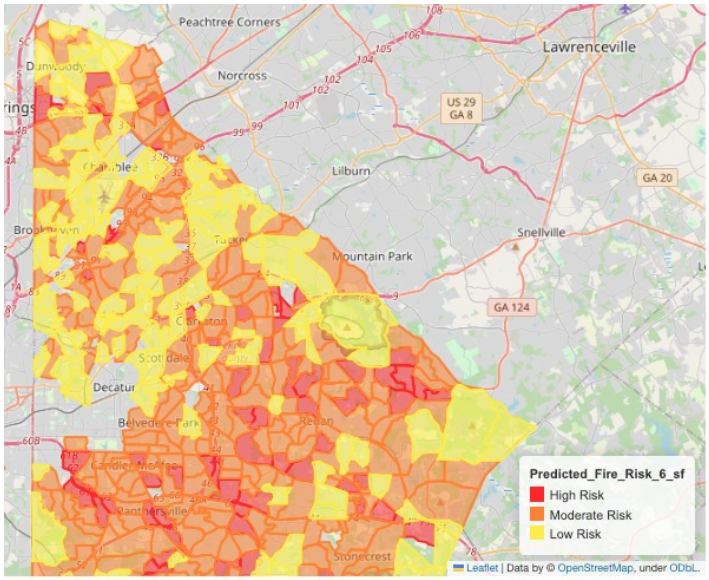
See A1. 6+ yrs. data, SF: BGID_9; Code: 'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb'

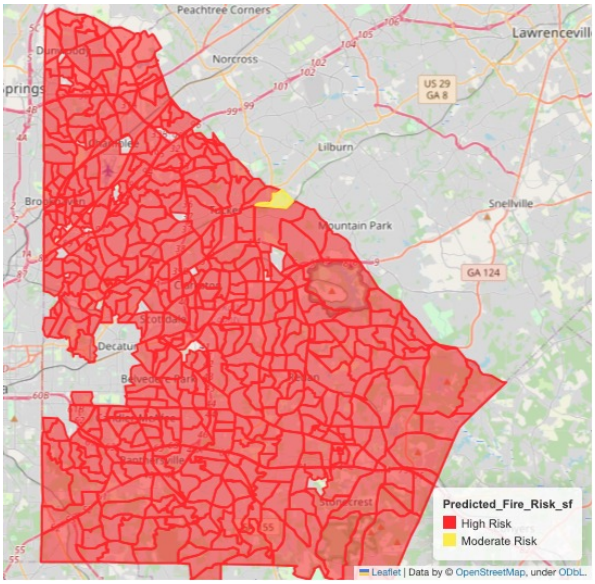
Comparison of Results of Single Family Approaches

Model fit to BGID 9

Model fit to BGID 6

Models fit to 6 + years of data





>> *Recommended*: A1. 6+ yrs. data, SF: BGID_9: Fire (0,1): (2084, 1308)

Accuracy of 75%, & good precision/recall/f1 for both Fire (0,1)

	BGID_9	BGID_6
High Risk:	**987**	**76**
Moderate Risk:	**1254**	**306**
Low Risk:	**1997**	**179**

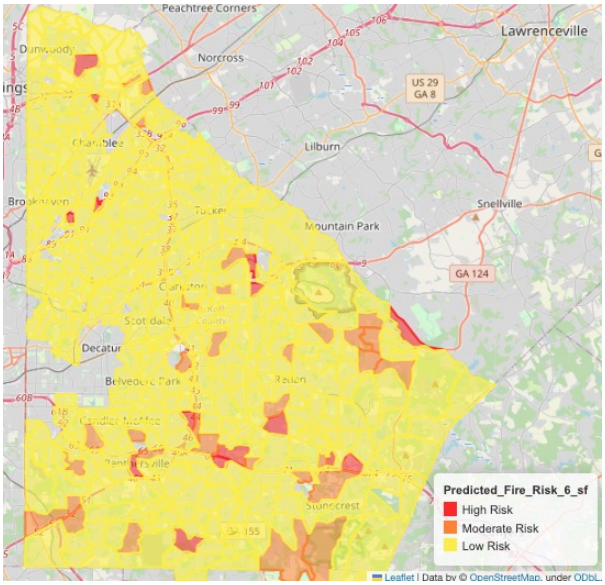
now: ph2_join_w_model.v5_MF-Copy4_LDA_6%2Byrs-Copy5b-Copy26.ipynb
was: ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb

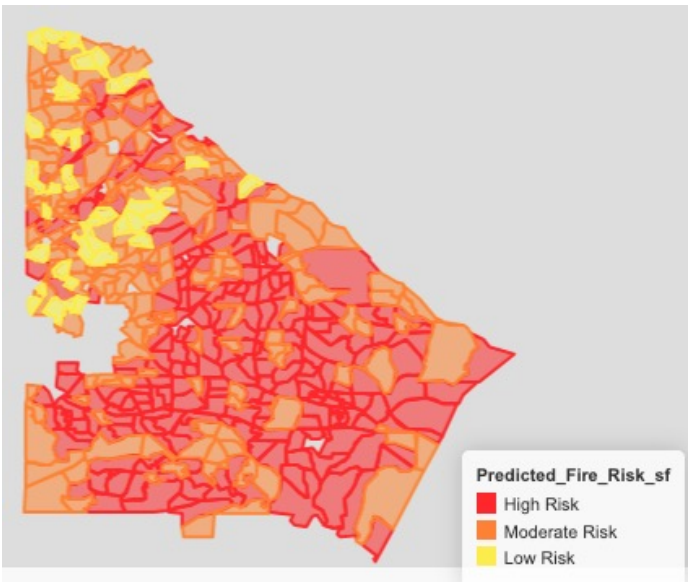
B1. 6+ yrs. data, SF: BGID_6:
Fire (0,1)= (3,460) Accuracy of 98%, but weak precision/recall/f1 for Fire (0)

	BGID_6
High Risk:	462
Moderate Risk:	1
Low Risk:	0

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy9.ipynb'

Models fit to < 1 year of data





C1. < 1 yr. data, SF: BGID_9:
Fire (0, 1)= (3123, 269) Accuracy of 91%, but weak precision/recall/f1 for Fire (1)

	BGID_9	BGID_6
High Risk:	987	76
Moderate Risk:	1254	306
Low Risk:	1997	179

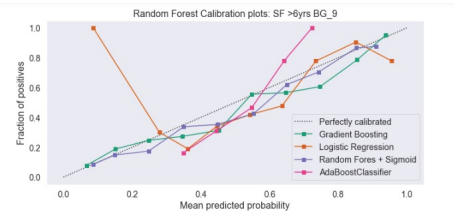
ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy21.ipynb

D1. < 1 yr. data, SF: BGID_6:
Fire (0,1)= (166,297) Accuracy of 74%, but weak precision/recall/f1 for Fire (0)

	BGID_6
High Risk:	244
Moderate Risk:	182
Low Risk:	37

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy24.ipynb'

Deep-Dive Comparison of Various Approaches: SF



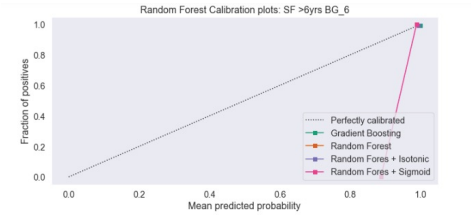

```
10 print(crsf_gbc)
```

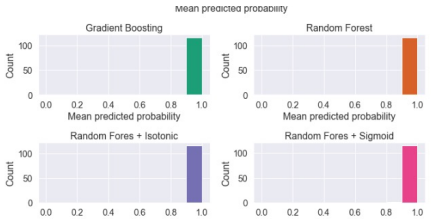
	precision	recall	f1-score	support
0	0.79	0.82	0.80	521
1	0.69	0.65	0.67	327
accuracy			0.75	848
macro avg	0.74	0.73	0.74	848
weighted avg	0.75	0.75	0.75	848

>>> A1. 6+ yrs. data, SF: BGID_9: Accuracy of 75%, & good precision/recall/f1 for both Fire (0,1)

ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy10.ipynb

Classifier SF >6yrs BG_9	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.17	0.52	0.80	0.69	0.65	0.67	0.75	0.47
Logistic Regression	0.23	0.66	0.66	0.51	0.53	0.52	0.62	0.20
Random Fores + Sigmoid	0.18	0.53	0.79	0.66	0.65	0.65	0.73	0.44
AdaBoostClassifier	0.20	0.58	0.80	0.66	0.65	0.65	0.73	0.44





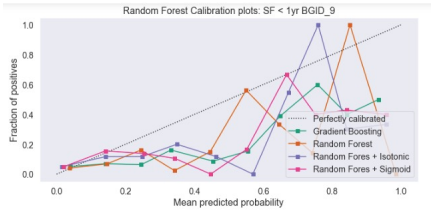

```
12 #rfcsf
```

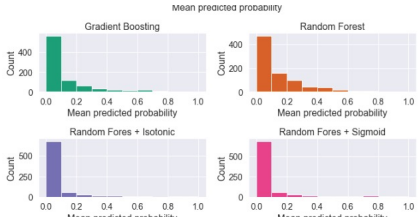
	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.99	1.00	1.00	115
accuracy			0.99	116
macro avg	0.50	0.50	0.50	116
weighted avg	0.98	0.99	0.99	116

B1. 6+ yrs. data, SF: BGID_6: Accuracy of 98%, but weak precision/recall/f1 for Fire (0)

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy9.ipynb'

Classifier SF >6yrs BG_6	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.02	0.23	0.50	0.99	0.99	0.99	0.98	-0.01
Random Forest	0.01	0.04	0.98	0.99	1.00	1.00	0.99	0.00
Random Fores + Isotonic	0.01	0.04	0.89	0.99	1.00	1.00	0.99	0.00
Random Fores + Sigmoid	0.01	0.04	1.00	0.99	1.00	1.00	0.99	0.00





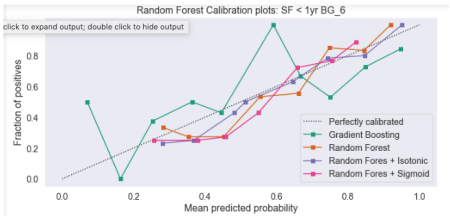

```
4 print(crsf)
```

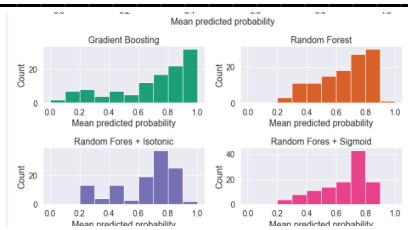
	precision	recall	f1-score	support
0	0.94	0.97	0.96	781
1	0.44	0.24	0.31	67
accuracy			0.92	848
macro avg	0.69	0.61	0.63	848
weighted avg	0.90	0.92	0.90	848

C1. < 1 yr. data, SF: BGID_9: Accuracy of 91%, but weak precision/recall/f1 for Fire (1)

ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy21.ipynb

Classifier SF < 1yr BGID_9	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.08	0.29	0.70	0.37	0.28	0.32	0.91	0.27
Random Forest	0.08	0.28	0.72	0.36	0.21	0.26	0.91	0.22
Random Fores + Isotonic	0.08	0.44	0.76	0.35	0.25	0.29	0.90	0.24
Random Fores + Sigmoid	0.08	0.29	0.76	0.40	0.27	0.32	0.91	0.28






```
11
12 #< 1 yr SF: BGID_6
```

	precision	recall	f1-score	support
0	0.66	0.45	0.54	42
1	0.74	0.86	0.80	74
accuracy			0.72	116
macro avg	0.70	0.66	0.67	116
weighted avg	0.71	0.72	0.70	116

D1. < 1 yr. data, SF: BGID_6: Accuracy of 74%, but weak precision/recall/f1 for Fire (0)

'ph2_join_w_model.v5_MF-Copy4_LDA_6+yrs-Copy5b-Copy24.ipynb'

Classifier SF < 1yr BG_6	Brier loss	Log loss	Roc auc	Precision	Recall	F1	Accuracy	Cohen kappa
Gradient Boosting	0.21	0.62	0.72	0.74	0.89	0.81	0.73	0.37
Random Forest	0.19	0.55	0.77	0.74	0.92	0.82	0.74	0.38
Random Fores + Isotonic	0.18	0.54	0.79	0.74	0.91	0.82	0.74	0.39
Random Fores + Sigmoid	0.19	0.56	0.76	0.72	0.91	0.80	0.72	0.32

MULTI FAMILY from FIRE II
Hypotheses regarding feature selection &
the Models' accuracy

Hypotheses 1: Using the full 6 plus years of data (1/16/2016 to 6/31/2022) will increase accuracy of the predictive model, despite the fact that relationships between some features and the dependent variable change before, during, and after the pandemic. This can be phrased more broadly as, "How to respond to the pandemic's impact on data?"

This is my initial reaction, supported by applying the model to different time periods. For this project, I would like to examine the data and algorithms involved in the correlation matrices, machine learning modes and/ or other relevant algorithms to verify whether or not my assumption is mathematically sound. If it appears my hypothesis is not mathematically sound, what steps should be taken to best address the impact of the pandemic on the data?'

In contrast, It has been suggested that a time period of less than one year, starting later in the pandemic (10/2021 to 6/31/2022), would be preferable, as it would fit the model to behaviors during that stage of the pandemic, perhaps reflecting a return to work by some, and excluding 2020, a period when so many stayed home. The thought was this approach would allow for a more realistic fit of the model to post-pandemic conditions. My concern with this is that due to the scarcity of fires, such a short time period is inadequate. Furthermore, as we see in Figure 1, below, the general movement continues in parallel before, during and after the pandemic.

Further exploration and discussion is warranted on this topic.

Overview: The change in relationships between Call 100s Fire and several other highly correlated call types is illustrated in *Figure 1*. Notice the trend lines for Call 100s compared to Call Types 300s, 600s, 400s, & 500s is higher in the years prior to the pandemic. This relationship changes around March 2020, with Call 100's trend line lower in relation to these call categories from approximately that date through the end date of the data, 6/31/2022. March 2020 coincides with when students, parents and many workers switching to working from home in DeKalb County and throughout the country.

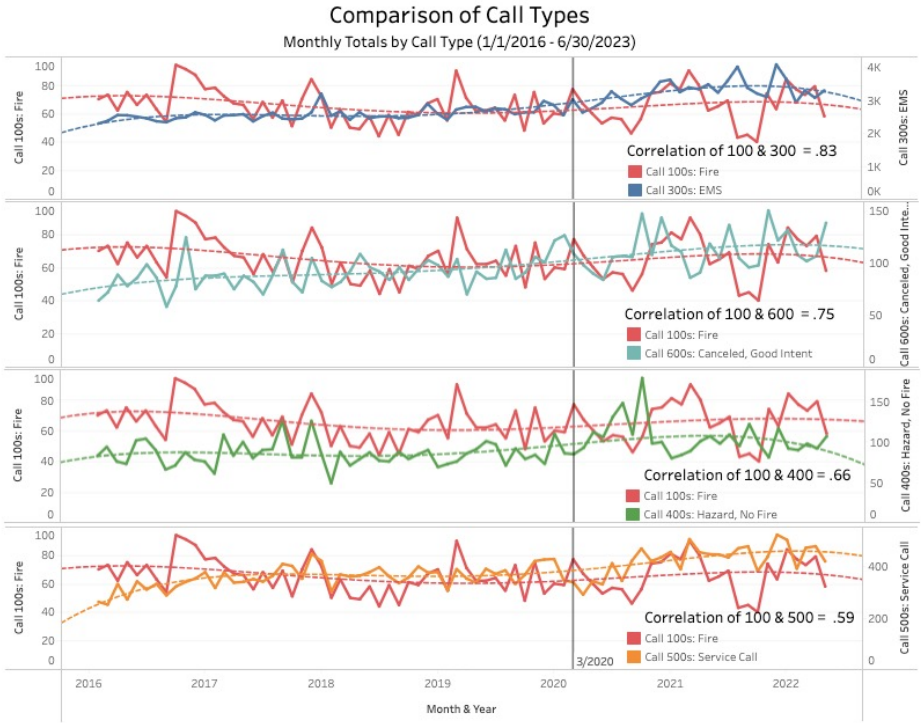


Figure 1: Comparison of multi-family call type trends, over time. Notice how the trendlines of "Call 100: Fire" versus other call types cross around 3/2020 (black vertical line), the beginning of the pandemic, as a large portion of society switches to working/studying remotely from their homes. Definitions of call categories from DCFR NFIRS data base: 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.

F. Findings & Deliverable.

Figure: A visualization and presentation of the model's results, including a deliverable: a table ranking DeKalb Single family properties' fire risk level, grouped at both the GEOIDs 12-and 15-digit , to be provided to DCFR for possible use in their Community Risk Reduction Assessment.

Deliverable: A prototype of an interactive app, with a map illustrating the probability of fire risk with the model fit to BGID 9, as described earlier.

Each dot is the center point of the 3392 BGID 9s. The probability of fire risk at the BGID 6 level was then determined by grouping these results by BGID 6 and calculating the mean.

Each of the 454 BDID 6s is colored to reflect this mean, as noted in the legend. Notice BGID 9 "212224002" is high risk, although its respective BGID 6 is yellow, or low risk. If scheduling permits, DCFR may target similar high risk BDGIS 9's within low risk BGID 6s. This app was created in Python, using the Streamlit library.

Fire Risk Single Family BGID 9: *dots*	
High Risk	945 black
Moderate Risk	1621 gray
Low Risk	1672 white

Fire Risk Single Family BGID 6: *colors*	
High Risk	59 red
Moderate Risk	390 orange
Low Risk	112 yellow