Predicting Fire Risk in NYC from Heating Complaints

Jesse Noppe-Brandon, Weijun Huang, Yiyang Zhou, Mike Zaharchenko

Table of contents

01

Hypothesis

02

Data Used

03

Data Preparation

04

Visualizations

05

Machine Learning

06

Conclusions

01

Hypothesis and objectives

Our Hypothesis

Buildings with a high volume of inadequate heating complaints are more likely to experience apartment fires in the near future. In the absence of sufficient heat, tenants may turn to unsafe alternatives like space heaters or ovens, which can increase risk.

Our Objective

Build a pipeline to analyze NYC Open Data datasets, train a machine learning model on them, and create a heatmap that shows grid-like regions that have the highest fire risk

02

Data Used

NYC® OpenData

~26M Records

Total records analyzed

Fire Incident Dispatch Data

~11M Records

Housing Maintenance Code Complaints and Problems

~15M records

03

Data Preparation

Stack

NYC
PLANNING
scikit
learn

Fire Incident Dispatch Data

First Processing Steps

- Drop the null value rows in the critical columns
- Filtered the dataset from the date range 01/01/2005-04/30/2025 to 01/01/2015-04/30/2025
  - Original: 11.1M rows (3.12GB) → After filter: ~6M rows (1.3GB)
- Filter the incident class that only contains the dwelling fire(6M to 126,200)
- Normalized addresses
- Geocached cross streets using the NYC Planning's Geoservice tool

Pipeline Optimization (pt. 1)

Leveraged Spark's partition mapping to process API calls concurrently and retrieve coordinates faster

RESULT

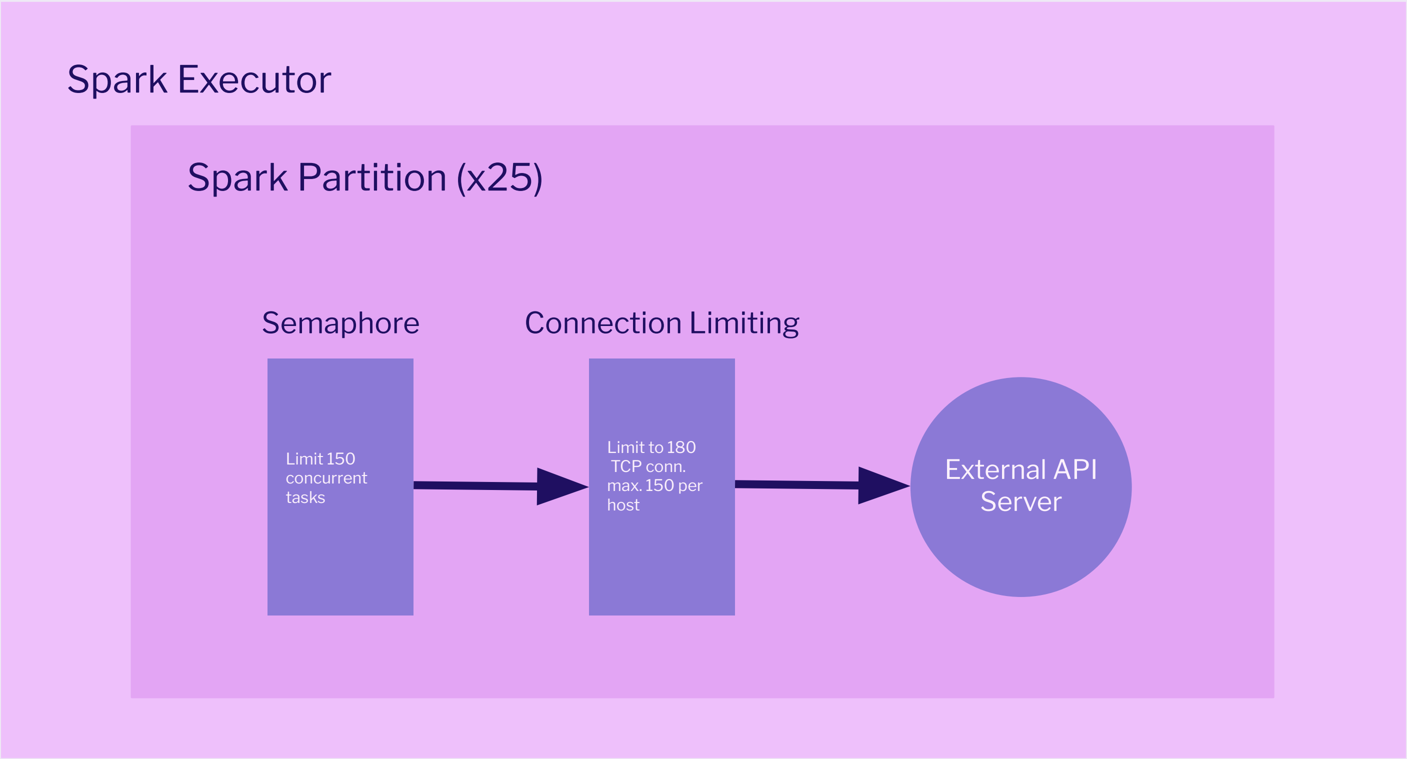
~6 hrs→ ~1.75 hr

to process entire data set

70.83% decrease

Pipeline Optimization (pt. 2)

Converted code to an asynchronous implementation using semaphores to protect critical sections and connection limiting to avoid overwhelming the API endpoint



Spark Executor

Spark Partition (x25)

Semaphore

Limit 150
concurrent
tasks

Connection Limiting

Limit to 180
TCP conn.
max. 150 per
host

External API
Server

RESULT

~1.75 hrs→ ~.5 hr

to process entire data set

71.43% decrease

Housing Maintenance Code Complaints and Problems

Preprocess House Complaint Data

- Drop the null value for the column in longitude and latitude
- Filtered the dataset from the date range 01/01/2005-04/30/2025
- to 01/01/2015-04/30/2025
- Only keep complaint with codes "NO HEAT AND NO HOT WATER" and "NO HEAT"
- Filtered out summer data because not representative

Match Complaints and Fire Incidents

Match Complaints and Fire Incidents based on time-region

Time:
Compare complaint data in the first 2 months and fire data in the next 2 months.

Geolocation:
To cover New York, we choose latitude from 40.490 to 40.920, longitude from -74.270 to -73.680.

New York is then divided into latitude-longitude grids with a resolution of 0.005 degrees.

Final Dataframe Schema

Region ID	A compound string of the latitude and longitude
Latitude	Latitude
Longitude	Longitude
Time Group ID	A year and whether it was in time group 0 (November/December complaints predicting January/February fires) or 1 (January/February complaints predicting March/April fires)
Total Complaints	Count of all complaints
Heating Complaints	Count of all heating-related complaints
Heating Complaints Percentage	Heating Complaints / Total Complaints
Fire Count	Count of fire incidents that occurred in the region
Previous Heating Complaints	Count of heating complaints in the previous time window of that region
Complaint Growth Ratio	The rate of change from the number of complaints in the previous time window to the current time window of that region.
Borough	The name of the borough the region is in

04

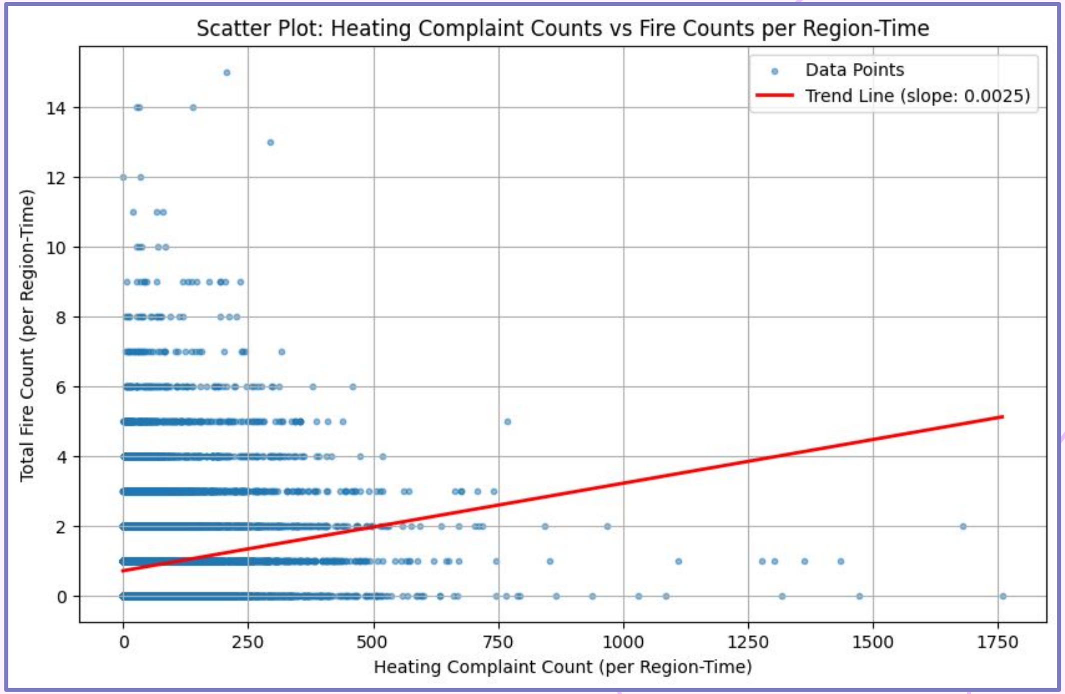
Visualizations

Heating Complaint Count vs. Total Fire Count

Plotted Heating Complaint Count vs. Total Fire Count across region-time windows.

The trend line shows a **weak positive correlation**

This suggests that heating complaints alone are **not a strong predictor of fires.**



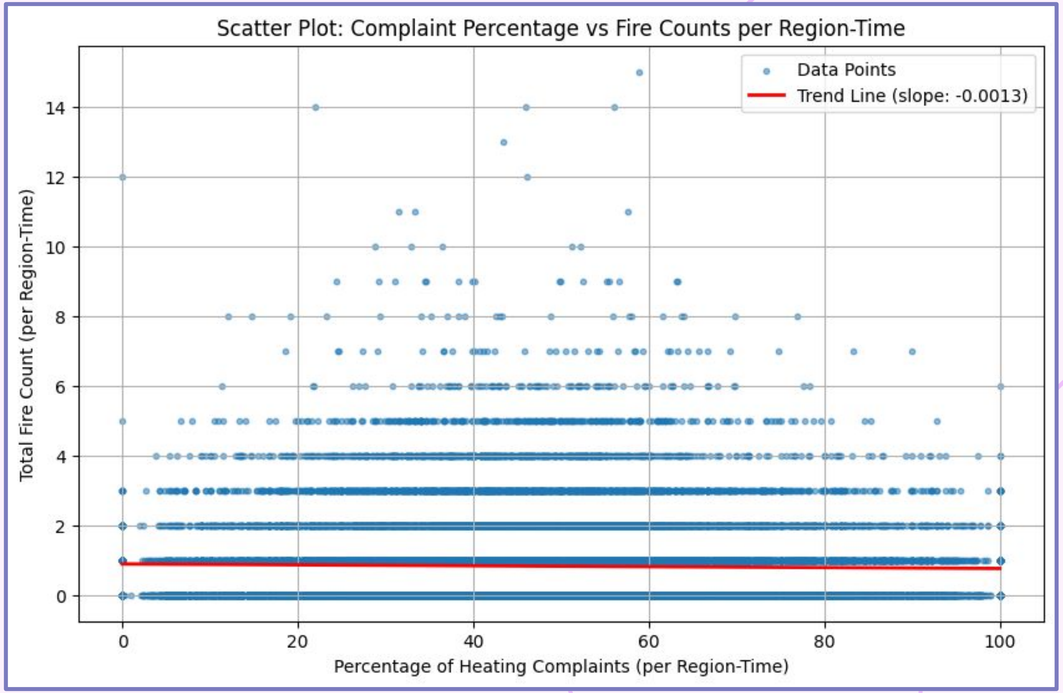
R-squared: 0.0256

% Heating Complaints vs. Total Fire Complaints

Plotted Percentage of Heating Complaints vs Total Fire Complaints

Wanted to account for the possibility of denser areas dominating the Number of Complaints

The resulting percentages were **heavily clustered** and showed **no meaningful trend** with fire occurrences.



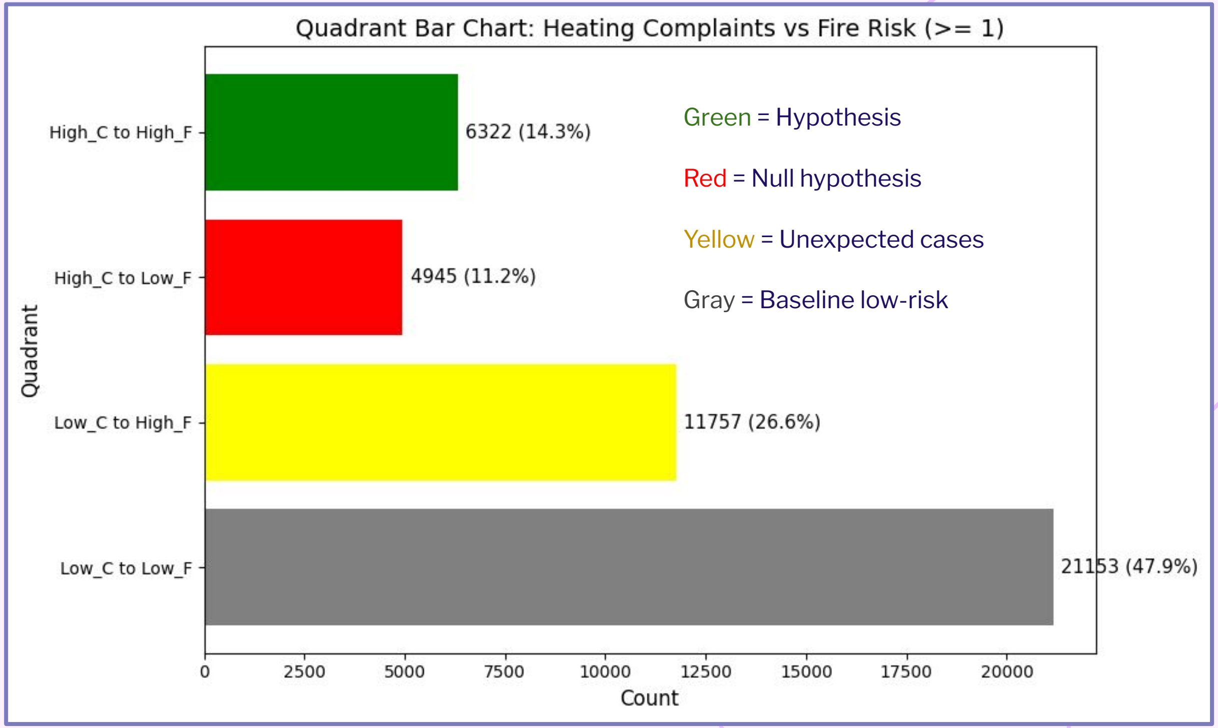
% of Heating Complaints = Heating Complaints / Total Complaints

Plotting our Hypothesis

We created this chart to plot our hypothesis.

High_C to High_F represents our *hypothesis*

High_C to Low_F represents our *null hypothesis*.



High_C: any region/time_window row that has a number of heating complaints above the 75th quartile.

High_F: any region/time_window row that has 1 or more fires.

Similar to the scatter plot, this quadrant chart shows our hypothesis (High_C → High_F) occurred in only **14.3% of cases**.

Most regions had 1 or fewer fires, meaning **high complaint counts often didn’t translate into high fire counts.**

Raising the fire threshold made our hypothesis quadrant (green) even smaller.

05

Machine Learning

Goal
Predict whether a fire will occur in a given region during the next 2-month period

Features Included

- Total complaints
- Heating complaints
- Percentage of heating complaints
- Previous period’s heating complaints
- Growth ratio in heating complaints
- Borough (one-hot encoded)

Why These Features?

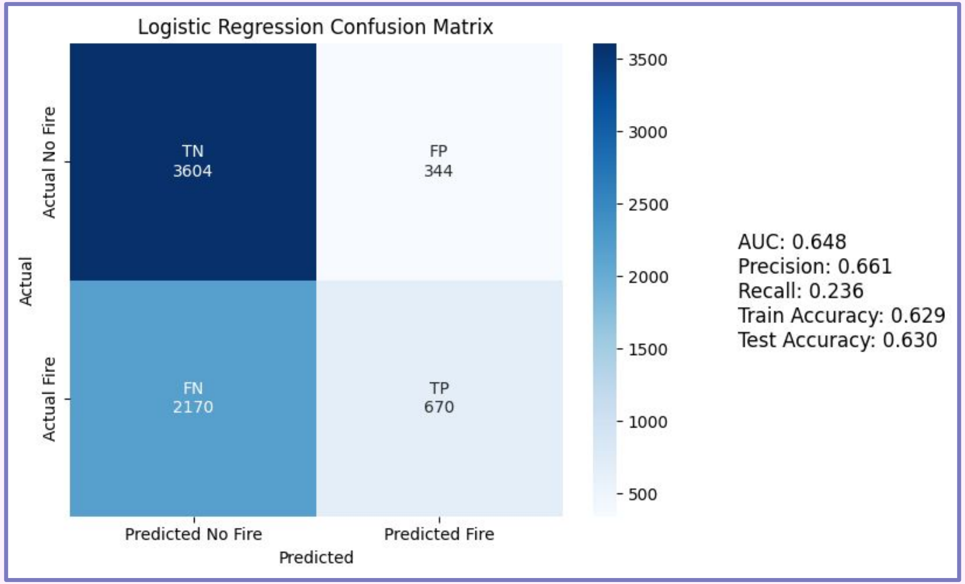
They capture **volume**, **change over time**, and **location** — the key dimensions we believed might influence fire risk.

Models Used

Logistic Regression

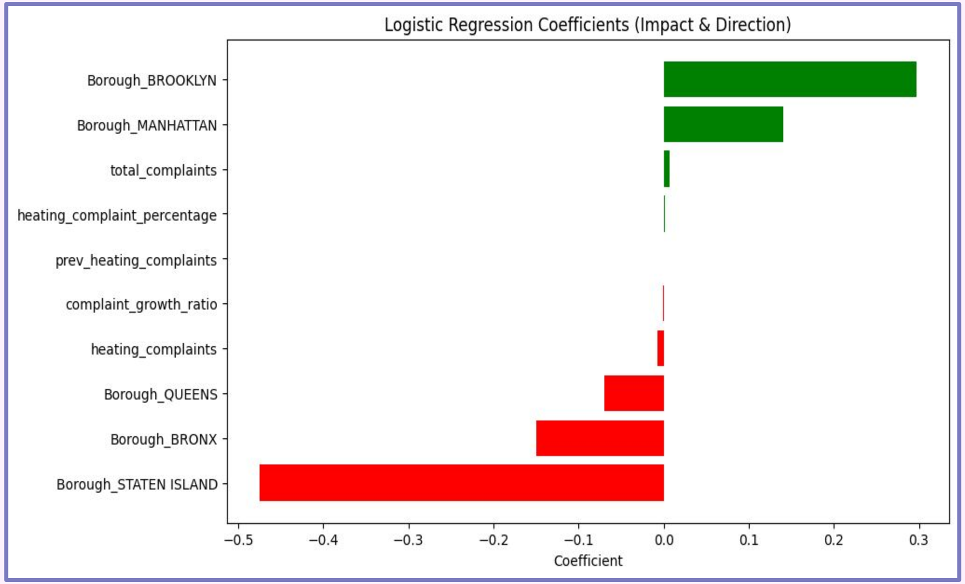
Chosen for interpretability

We can understand which features increase or decrease fire risk



Confusion Matrix

- Test accuracy of 63%
- AUC of 0.648
- 23.6% recall; very low
- Decent precision score; not surprising given how much more common not having a fire is than having one

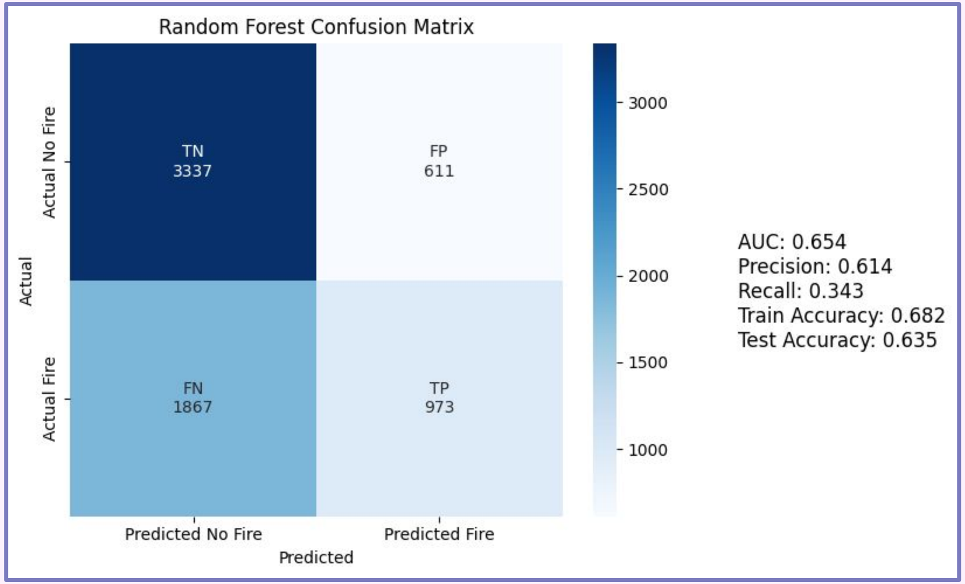


Coefficients

- Borough played a strong role in model's prediction, which makes sense given varying population densities
- Heating complaints had a slightly negative coefficient
- Total complaints showed a modest positive effect
- Tested a model without borough as well; achieved similar results

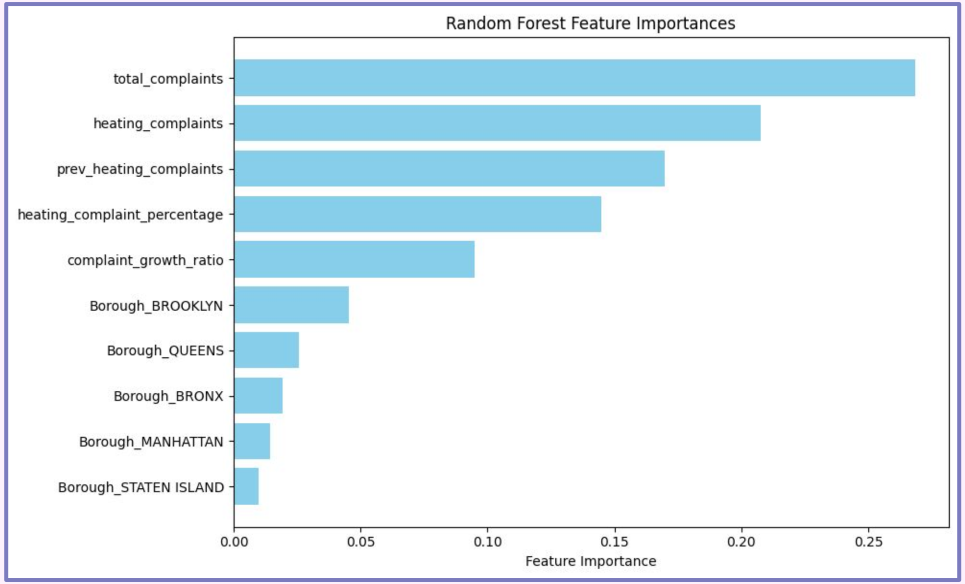
Random Forest

A more complex, tree-based model to potentially capture nonlinear relationships



Confusion Matrix

- Test accuracy and AUC nearly identical to Logistic Regression model
- 34.3% recall; slightly higher, but still too low for effective risk detection
- Very similar precision score (61.4%)



Feature Importance

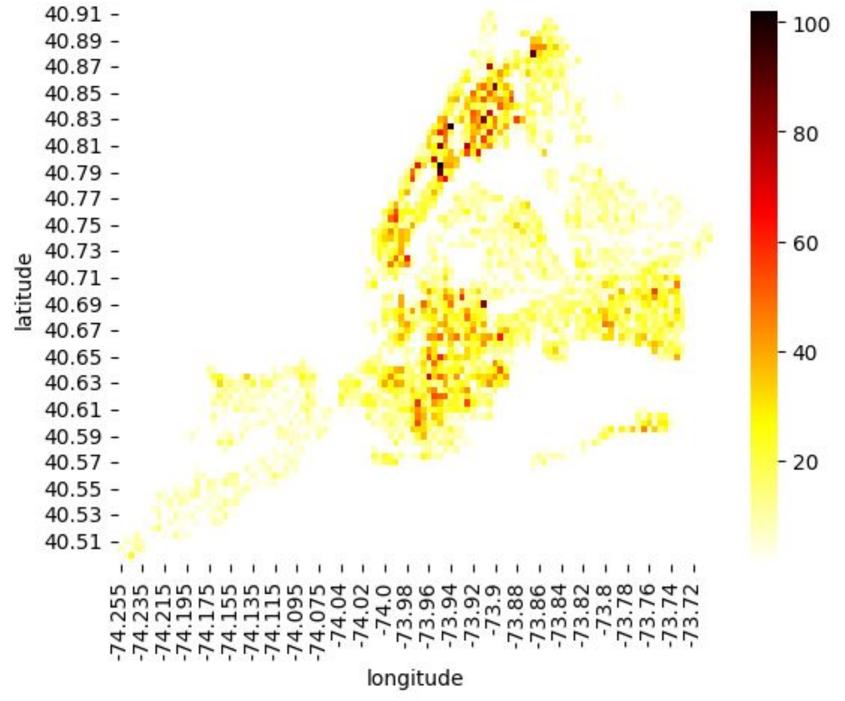
- Borough features played a smaller role here
- Total complaints was also the highest predictor of the numerical categories
- Heating complaints were in second place
- Both models performed similarly, but RF placed a greater emphasis on complaint volume over location

06

Heatmap & Conclusions

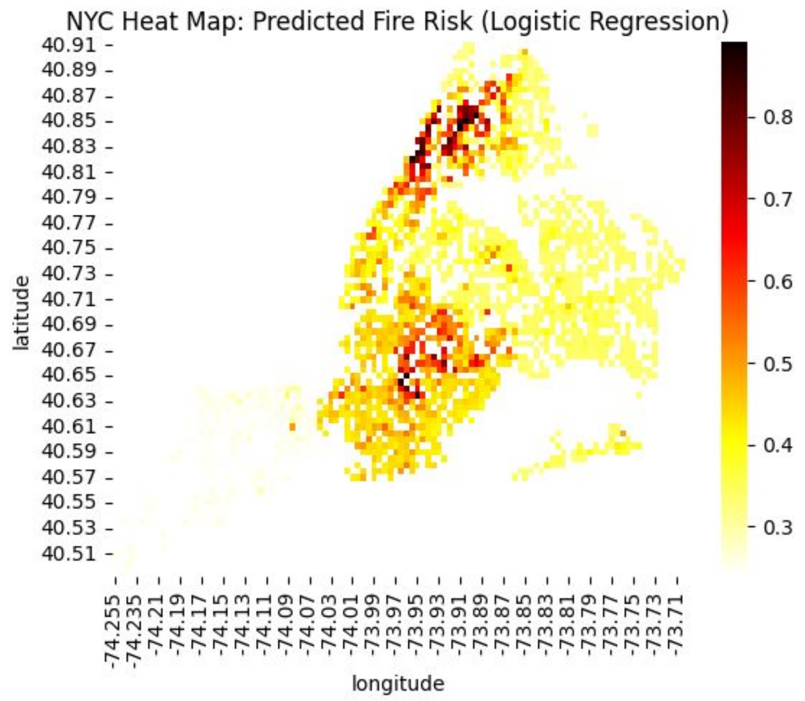
Heat Map: Fire Count

- Manhattan and Brooklyn had the highest fire counts.
- Given the relative population density of those two boroughs, this makes sense



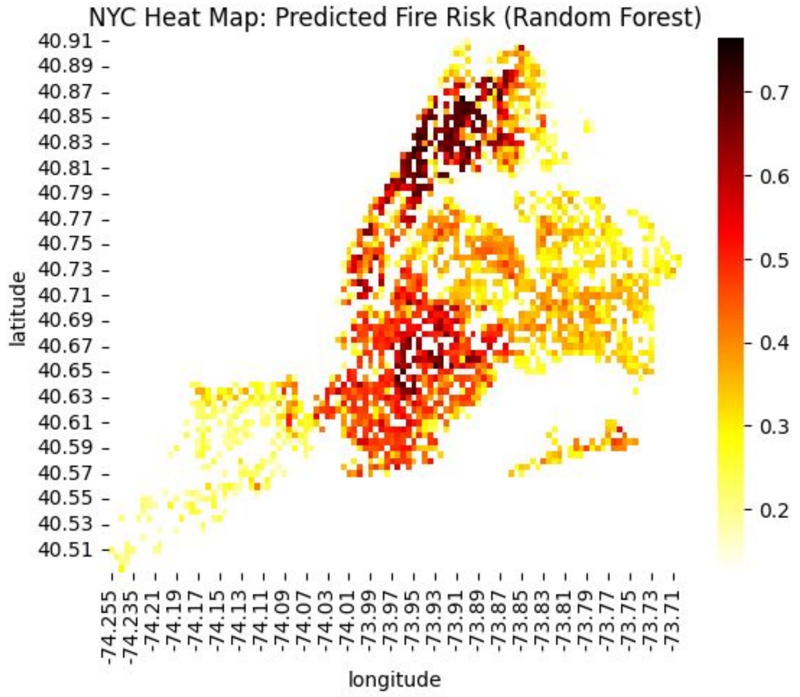
Heat Map: Predicted Fire Risk (Logistic Regression)

- Staten Island is nearly invisible
- Spatial patterns similar to the fire count heatmap are especially visible in the logistic regression map
- Heatmap remained virtually unchanged when borough info was excluded from the model
- Indicates that location patterns correlated to other features



Heat Map: Predicted Fire Risk (Random Forest)

- Both models show spatial patterns similar to the fire count heatmap
- Random forest produces more dispersed hot spots as compared to logistic regression
- Staten Island is more visible in the random forest heatmap



Conclusion

Overall, all of our data exploration showed that the number of Heating Complaints is inadequate to predict fires by itself. In future work, we would try adding more features such as demographic information and building data to see if they could increase our predictive power. This could still support our original goal of identifying areas at greater fire risk due to neglect.

Interestingly, the heat maps created from the models showed a somewhat similar pattern to the fire count heat map. However, this mostly correlated with denser areas like Manhattan and Brooklyn, which also had more fires overall.

If more features led to more robust predictions, this project could be extended to a streaming pipeline. Housing Complaint and Fire Incident data is updated regularly (new data arrived while we were working on this project), and a pipeline could continually process this incoming data and use it to improve predictions.

While our hypothesis didn't hold, heating complaints still seem to be a small but relevant piece of the puzzle. This landlord neglect affects tenants' safety and well-being, and whether or not it directly causes fires, better care of buildings and tenants leads to a safer city.

Thank You