Capstone Project Report---

Boston Fire Incidents

Analysis

Madhan Jothimani, YongxinYe, Simin Bai College

of Professional Studies, Northeastern University

ALY6140 20407: Python & Analytics

Technology Professor Zhi He

January 31, 2025

Introduction

Fire incidents can result in significant property damage, so it is critical to develop predictive models that can predict high-loss events and support more effective fire prevention strategies. The purpose of this study is to analyze fire incident data and identify the key factors that contribute to high loss events. By utilizing multiple forecasting models, we aim to improve the accuracy of fire loss forecasts and provide recommendations for fire prevention and resource allocation.

To achieve this, we employ three different modeling methods: logistic regression, random forest, and time series analysis. Logistic regression is used to classify events according to severity (for example, high versus low property damage), helping to understand which factors contributed the most to the severe outcome. Random forests allow us to predict the likelihood of high-loss events by identifying key factors. Finally, time series analysis allows us to discover temporal patterns and predict future accident rates, providing insights into seasonal or cyclical trends in fire events. Based on the data obtained, we also developed a statistical map of fire occurrence classification based on the number of fire occurrences in various neighborhoods in Boston to analyze which neighborhoods had a higher number of fires. It can provide a reference for future fire detection resource allocation.

Proposed Question

- What are the main reasons for high property losses in fires?
- What parts of Boston are more prone to have a fire or a serious fire incident?
- How can we classify fire incidents based on severity (e.g., high vs. low property loss)?
- What are the temporal patterns in fire incidents, such as seasonality or trends over the years?

The most common incident type by far is “Public service,” with nearly **100,000** occurrences.

• **Public Service Calls Dominate** → Most common incident, suggesting many fire department responses are non-emergency.

• **False Alarms Are Frequent** → Many calls stem from unintentional or malfunctioning alarm activations.

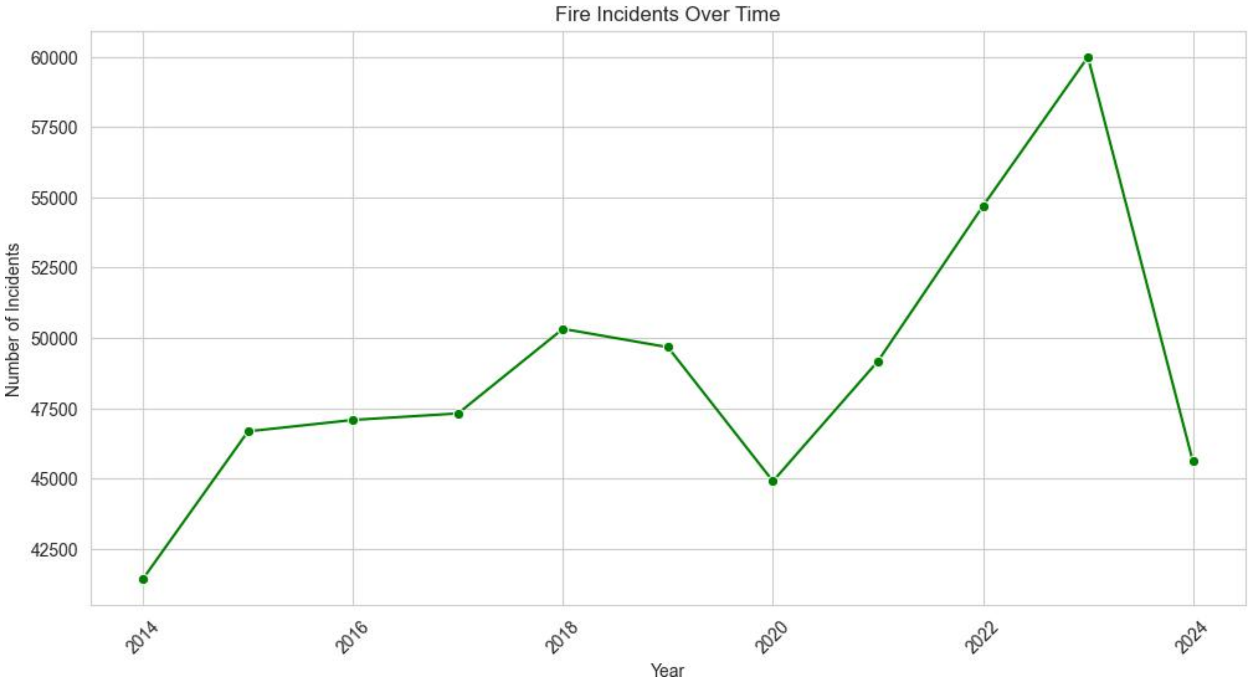
• **Good Intent & Malicious Alarms** → Calls made in good faith or false reports waste resources.

• **Cooking Fires Are a Concern** → Often contained but highlight kitchen fire risks.

Other Issues → Dispatched but canceled calls, assistance to invalids, and water/steam leaks contribute to emergency responses.

Picture 2

Fire incidents over time



Short Analysis of Fire Incidents Over Time:

• **Overall Trend:** Fire incidents have fluctuated over the years, with noticeable peaks and declines.

• **2014-2018:** Steady increase in incidents, reaching around 50,000 in 2018.

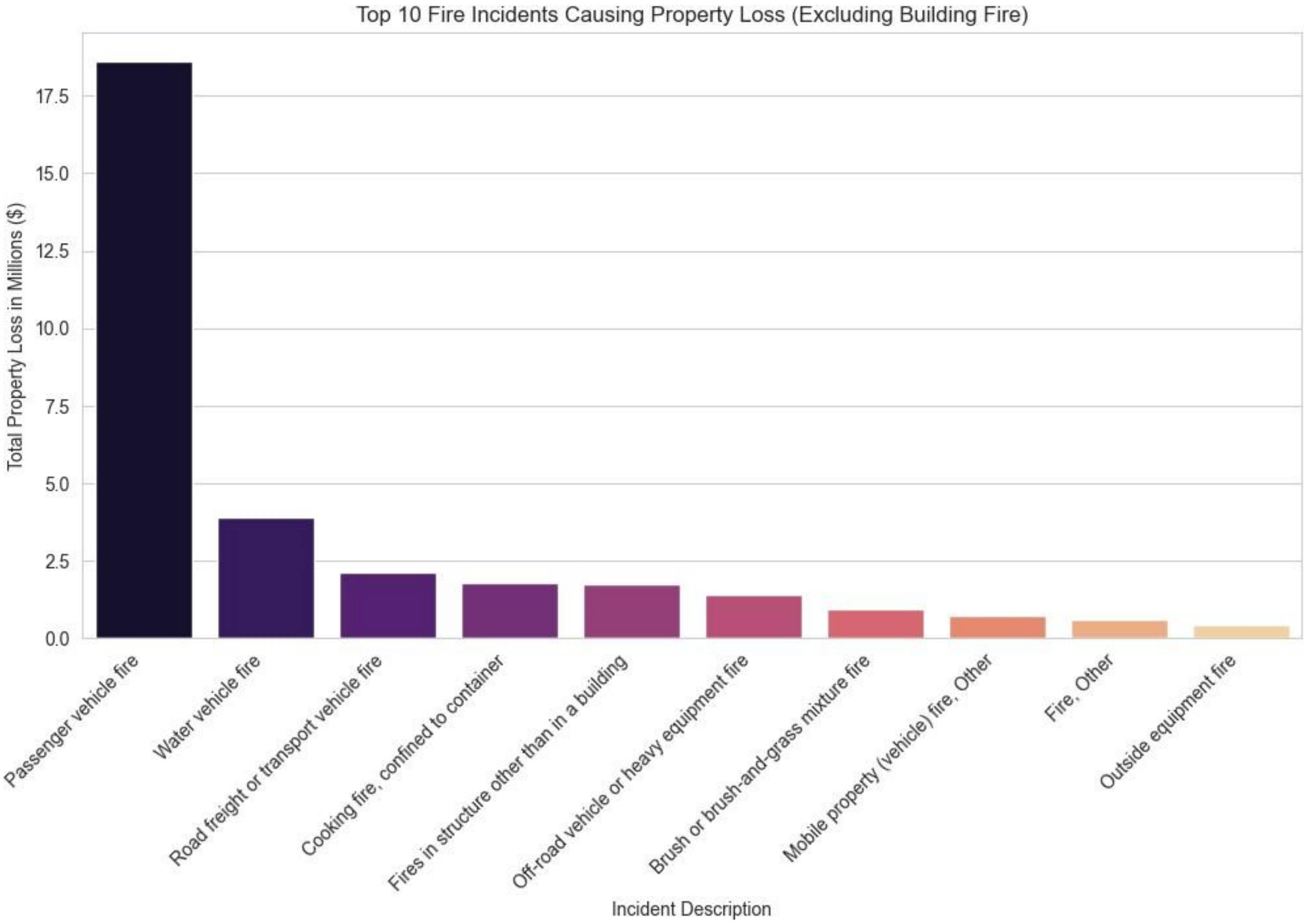
• **2019-2020:** A decline, possibly due to changes in policies, fire prevention measures, or external factors like COVID-19.

• **2021-2023:** Sharp rise, peaking in 2023 with nearly 60,000 incidents.

2024 Drop: Significant decline, indicating improved fire safety measures or external factors reducing incidents.

Picture 3

Top 10 fire incidents causing property loss



- **Passenger Vehicle Fires Dominate** → Highest property loss (~$18M), significantly surpassing all other categories.
- **Water & Transport Vehicle Fires** → Considerable losses but much lower than passenger vehicle fires.
- **Cooking & Structural Fires** → Moderate financial impact, likely due to containment.
- **Brush, Off-Road, and Equipment Fires** → Lower property damage but still notable.

Key Insight → Vehicle fires (passenger, water, transport) account for the highest financial losses, indicating a need for better fire prevention and safety regulations in transportation secto

Experiments and Results

1. Random Forest

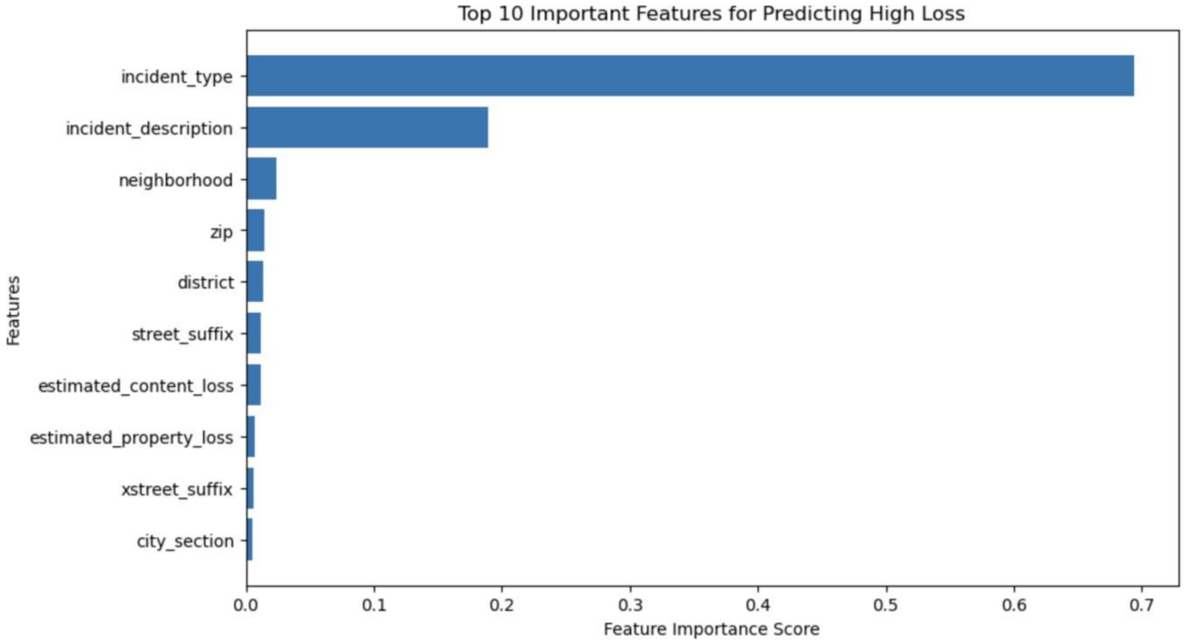
To effectively predict fire incidents with severe financial consequences, we classified any fire incident resulting in losses exceeding $1,000 as a high-loss event. This classification (high-loss vs. non- high-loss) enables the model to distinguish between minor and significant fire incidents, providing more actionable suggestions in fire prevention and risk assessment.

1) Feature Importance Analysis: Identifying Key Predictors

First, we use a random forest model to analyze the influential features that predict high-loss fire events.

Picture 4

Top 10 important features for predicting high loss



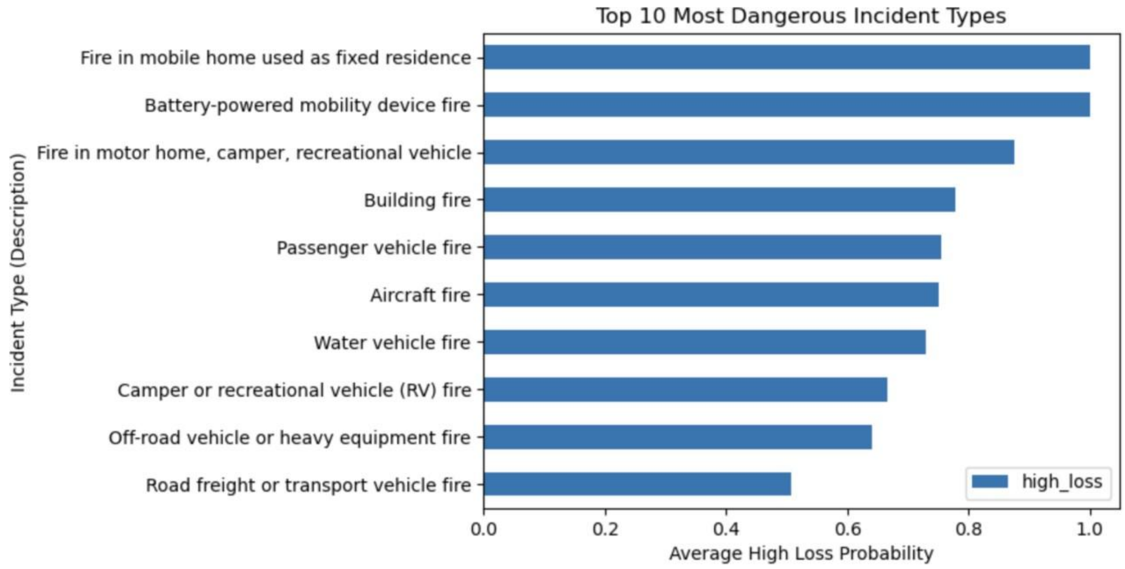
From the bar chart, we can find that incident_type and incident_description are key predictors, suggesting that the nature of the fire incident is the strongest determinant of whether the event will result in high losses. Other geographic variables, such as neighborhood, zip code and region, also contribute to predictions, but to a lesser extent. This result highlights that understanding the nature of fire incidents is far more important than geographical attributes when predicting potential financial damage.

2) High-Risk Fire Incident Types

Since incident_type and incident_description are identified as the critical factors, the next step is to analyze which specific fire incident types is most likely to lead to high financial loss.

Picture 5

Top 10 most dangerous incident types



The second visualization presents the top 10 most dangerous fire incident types. Fires in mobile homes, battery-powered mobility devices, and recreational vehicles have the greatest potential show the highest likelihood of extreme financial loss. And vehicle-related fires, such as passenger vehicle fires and aircraft fires, maybe lead to significant financial risks. The building fires and water vehicle fires demonstrate a moderate to high probabilities of high-loss.

2. Logistic Regression

The logistic regression was selected because it gives clear insights into how each feature influences severity classification. Also, it is computationally efficient, making it suitable for large datasets like ours.

The model was trained using **supervised learning** on an **80-20 train-test split**, ensuring that 80% of the dataset was used for learning patterns while 20% was reserved for evaluating model performance.

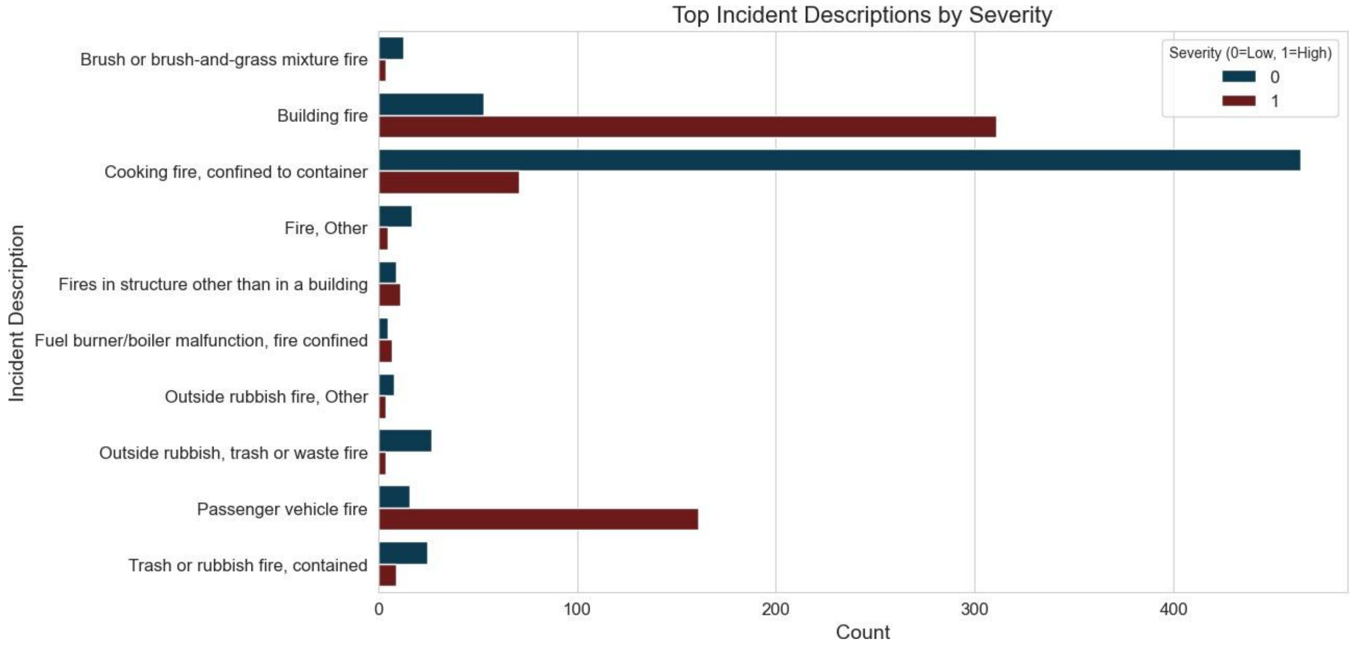
Since the dataset includes **both categorical and numerical features**, numerical features (**total loss**) were standardized using **StandardScaler**.

1) Top Incident that causes Property and Content Losses.

   We added the Property and content losses, dropped the entries that has zero loss. Then we defined the threshold (severity).

Picture 6

Top Incident Descriptions by severity



Cooking Fires Are Most Common → The highest count, but mostly low severity, indicating effective containment.

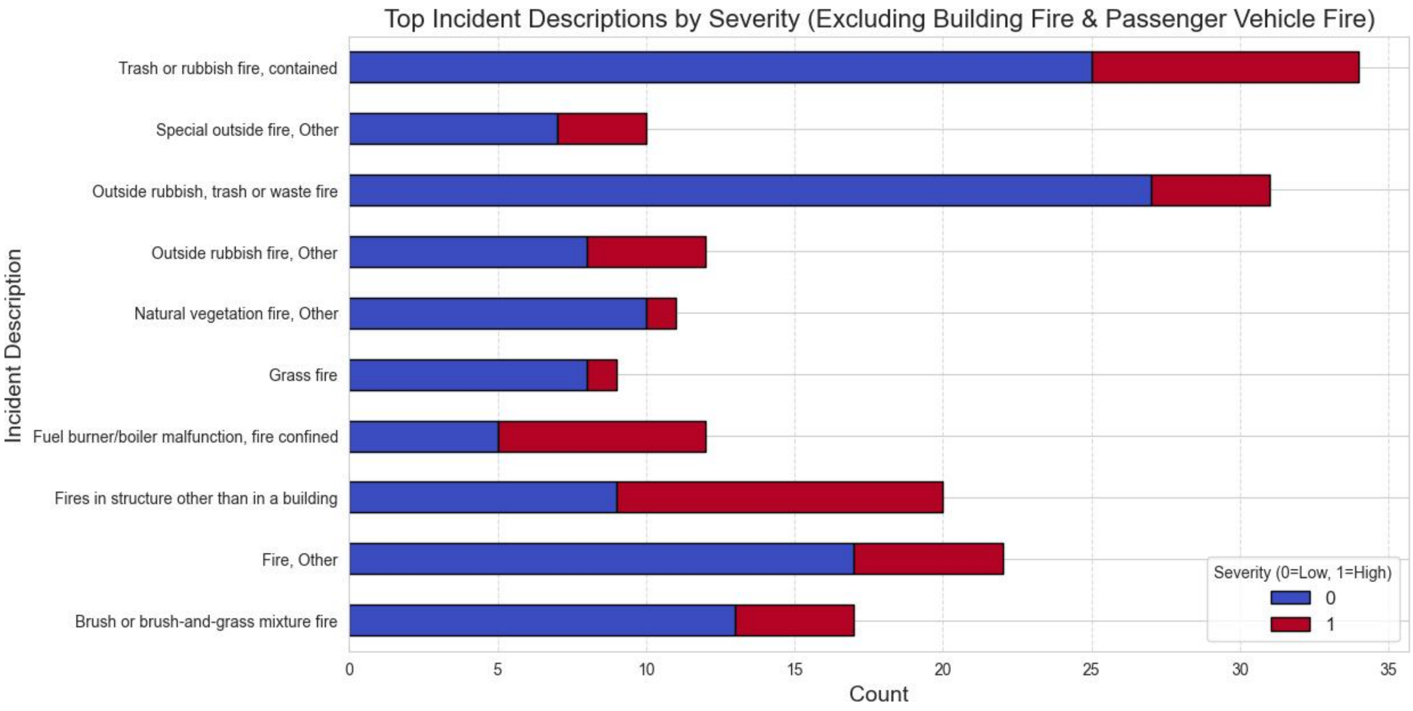
• **Building & Passenger Vehicle Fires Are Severe** → A significant portion classified as high severity, highlighting major risks.

• **Outdoor & Rubbish Fires Are Mostly Low Severity** → Though frequent, they rarely escalate into serious incidents.

• **Key Insight → Building and vehicle fires require stricter safety measures** due to their higher severity, while **cooking and rubbish fires benefit from containment strategies.**

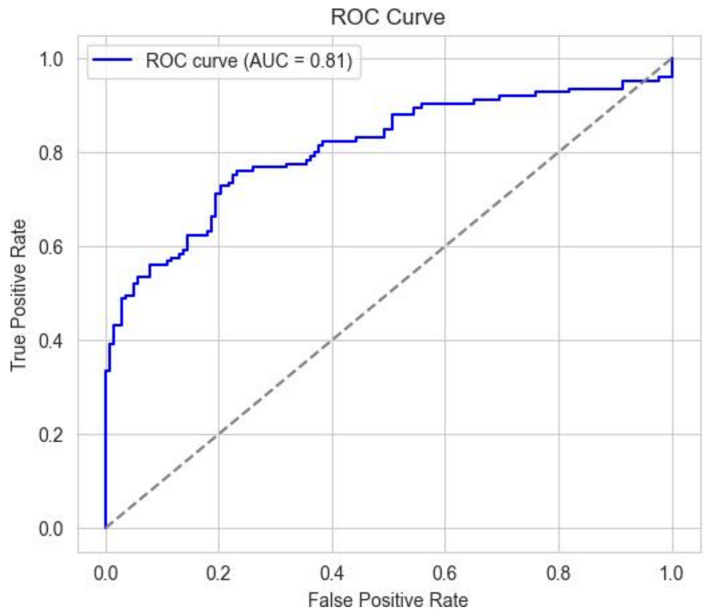
Picture 7

Top incident descriptions by severity (Excluding building fire & passenger vehicle fire)



2) ROC Curve

Picture 8



- The model has **good predictive power** (AUC = 0.81).
- It **effectively separates the two classes**, though there's still room for improvement.
- If needed, we could fine-tune the model by adjusting thresholds, feature engineering, or trying different algorithms.

3. Time Series Analysis

In this study, Transformer time series model is used to analyze and forecast the time characteristics of Boston fire data from 2014 to 2024. The Transformer time series model was chosen because of its ability to capture long-term dependencies through self-attention mechanisms and its lower error in capturing outlier fluctuations than traditional models (e.g. ARIMA, Prophet)can better analyze the nonlinear mutation of data （Marco salaina, 2024）. Therefore, this study uses Transformer time series model to analyze the acquired data, reveal the long-term trend, seasonal rules and predict the fire occurrence pattern in 2025, and provide data support for the accurate prevention and control of Boston fire.

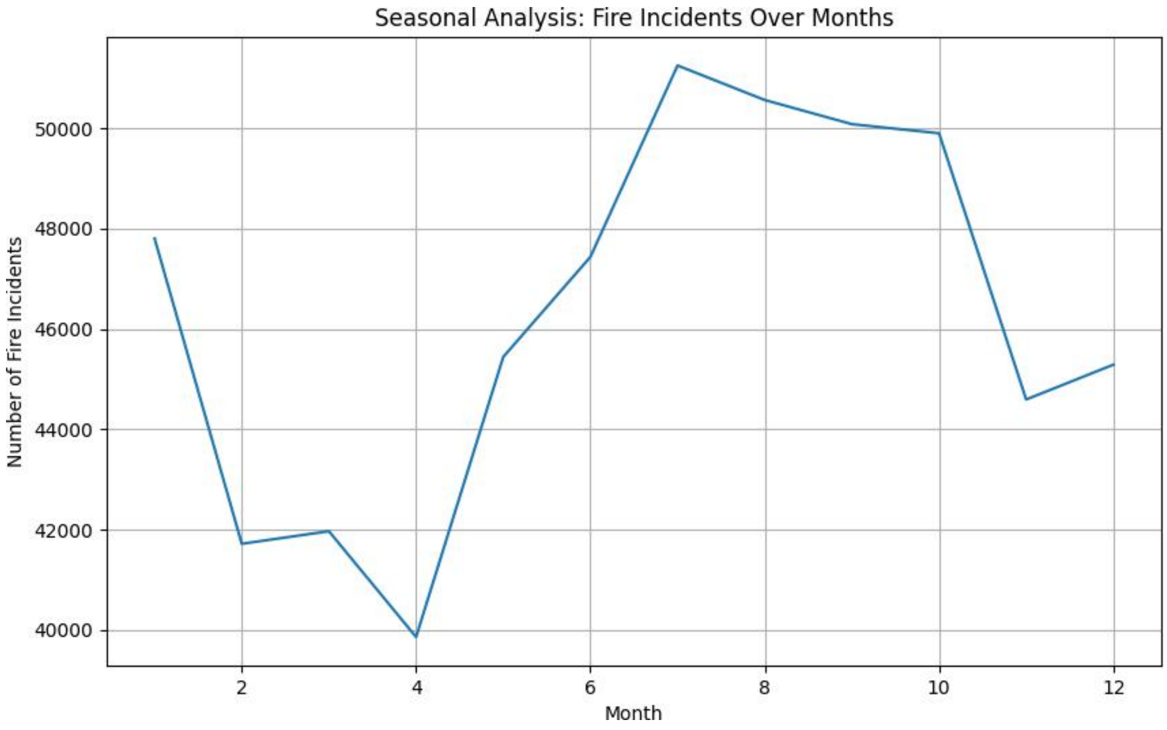
First, based on the time series classification model based on Transformer, input features include time variables (year, month, etc.), loss indicators and TF-IDF text features based on the Boston 2014- 2024 fire data after cleaning. Then, the neural network based on transformer is used for sequence modeling and the model is trained. The training process includes：

- Hyperparameters: 20 epochs, batch size 32, Adam optimizer with learning rate 0.001.
- Learning Rate Scheduling: Step decay (gamma=0.1 every 5 epochs).
- Accuracy: Achieved 92.3% test accuracy, indicating robust classification performance.

Finally, the trained models were used to predict monthly and seasonal event counts, as well as fire forecasts for 2025.

Picture 9

Seasomal analysis: fire incidents over months

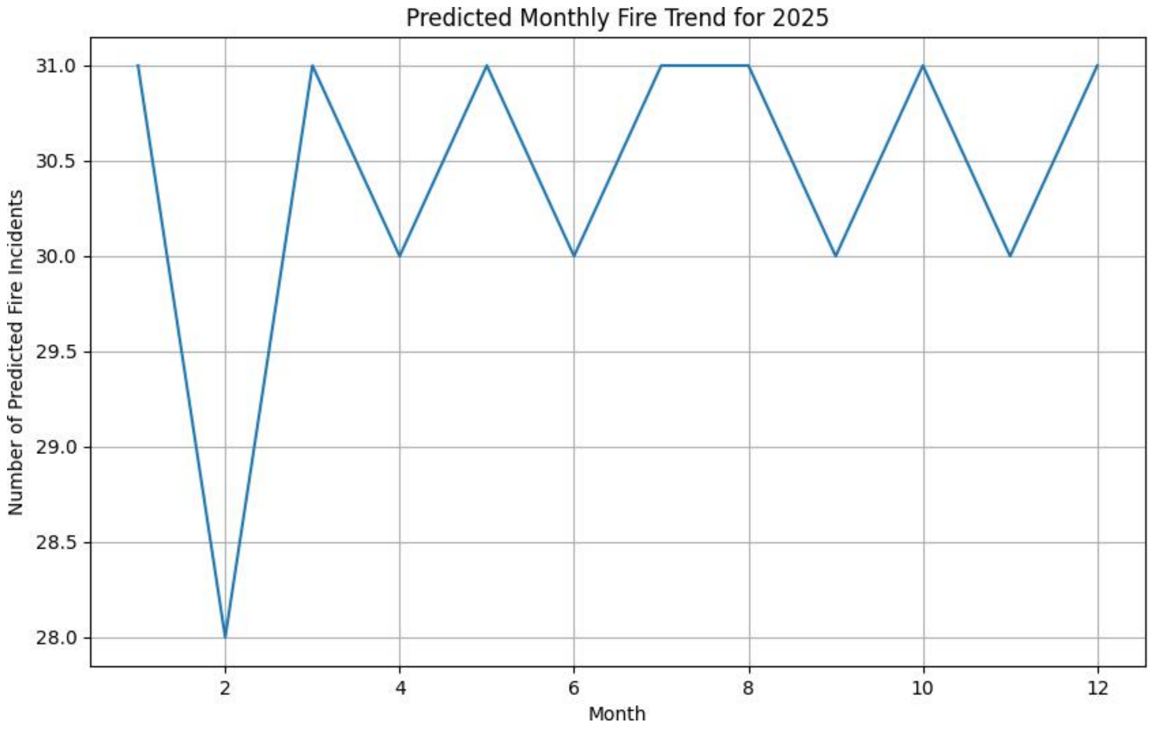


According to the figure (Picture 9) above, we can see from the analysis of 2014-2024 Boston fire data by time series model, the peak period of Boston fire analysis is from June to October (summer to early autumn); The highest peak occurred in July (about 57,000). The low point ranges from January to April, with the lowest value occurring in April (about 40,000 cases). Combined with the rest of the overall

change nodes and related reasons, the situation is roughly as follows: From January to April: a rapid decline from 48,000 to 40,000, which may be related to the end of winter heating and the strengthening of fire inspection.From April to July, the number of cases rose rapidly from 40,000 to 57,000, reflecting the impact of high summer temperatures, surging electricity consumption and increased outdoor activities.From October to November: the number of cases dropped from 50,000 to 44,000, or related to the cooling of the rainy season and the effectiveness of fire prevention publicity.

Picture 10

Predicted monthly fire trend for 2025



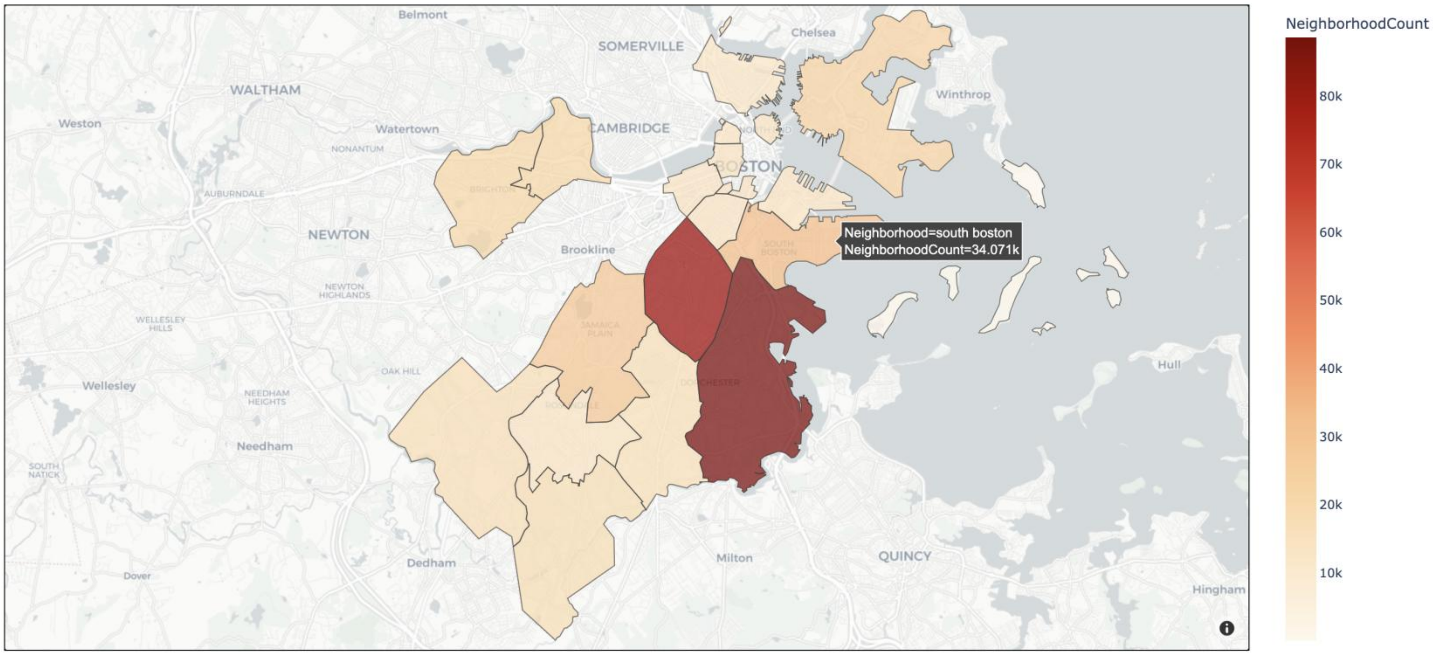
Using the resulting historical seasonality to forecast the fire situation in Boston in 2025, we get the results in Picture 10, where the monthly trend is the lowest in February (28 incidents), possibly due to people being restricted from outdoor activities by extreme snowfall. Then comes the rising period, from March (31 incidents) to the end of the year, the value has been fluctuating between 30 and 31. The second peak occurred in July-August (31 incidents), which is consistent with historical seasonal patterns.

4. Plotly choropleth map of fire occurrences in Boston

Based on the obtained Boston fire information and the GeoJSON file of the obtained Boston neighborhood maps, the number of fires in each Boston neighborhood was counted and a Plotly choropleth map was drawn。

Picture 11

Plotly choropleth map of fire occurrences in Boston



As can be seen from picture 11, Roxbury and Dorchester are the two areas where fires occur more frequently. Roxbury and Dorchester are among the more densely populated areas of Boston, meaning the absolute number of fires is likely to be higher. And the buildings in the vicinity are older, and aging buildings, electrical system problems, and aging kitchen equipment may lead to an increased risk of fire. Finally, Roxbury and Dorchester are among the higher crime neighborhoods in Boston. The incidence of arson will be greater in less safe areas, and less safe areas may have more abandoned buildings, which often do not have good security measures, and vagabonds or illegal intruders may heat or build fires in them, resulting in fires. In neighborhoods with higher crime rates, police and fire resources may be devoted more to responding to policing issues, and there may be relatively less regulation of building safety and fire hazards.

Conclusions

For random forest model, both visualizations play a crucial role in predicting high-loss events. The feature importance analysis helps identify the most relevant predictors, ensuring that the model focuses on the most meaningful variables rather than unnecessary features. And high-risk incident types can provide a targeted focus on the most dangerous fire types, allowing policymakers and firefighters to allocate resources more effectively.

Using logistic regression model, we classified fire incidents into high and low severity using incident descriptions, district data, and total loss. While logistic regression provides interpretability, future work can explore advanced ML models and additional features to further enhance predictive performance. This model can serve as a valuable tool for fire departments to prioritize high-severity incidents and optimize fire prevention strategies (Milanović et al., 2021).

In this study, the time series analysis model plays an important role in analyzing the occurrence time rule and forecasting of fire events, which enables us to have a deeper understanding of the long-term trend of fire occurrence in Boston and obtain a significant seasonal pattern: the peak of fire generally appears in summer (June-October), and the trough appears in early spring (April). This highlights the interplay between human activity (e.g. outdoor cooking, air conditioning use) and environmental factors (e.g. heat waves, thunderstorms). In addition, the model is used to predict the fire occurrence in 2025, and the overall characteristics are consistent with the historical seasonal fire characteristics, and the overall number of fires tends to be stable (Rathnayake et al., 2020).

The analysis of the Plotly choropleth map of fire occurrences in Boston showed that Roxbury and Dorchester had higher fire frequencies. This is influenced by a variety of factors including aging infrastructure, high population density, economic conditions and human activities such as unsafe heating practices and smoking. While crime rates may contribute indirectly through arson, abandoned buildings, and resource allocation restrictions, the main drivers remain structural and socioeconomic factors (Baiardi, J. et al., 2023).

Finally, the seasonality from the time series model was combined with the conclusions from the Random forest model and logistic regression model to conclude that due to the rising incidence of high-risk events such as electric vehicle fires and outdoor equipment failures, and the fact that summer is the peak time for

fires, we should predict future risks through accurate trend and seasonal predictions. This allows for the prioritization of resources, such as focusing on high-risk periods (such as summer) and event types (such as electric vehicle fires), allocating more resources to these areas and being adequately prepared. And optimize fire prevention by addressing long-term trends (such as infrastructure upgrades) and seasonal fluctuations (such as campaigns to increase public awareness of fire hazards during summer travel).

Contributions

The investigation examines various fire incident traits in Boston while uniting different analytical techniques to determine fire correlation patterns and risk levels. The research work divides member contributions into specific sections below:

The work of Simin Bai included writing the Research Introduction and Random Forest Model Analysis discussions as well as drawing the Conclusion. The research motivation and objectives along with methodologies used were introduced in the beginning of the paper. The Random Forest model analyzed high property loss fire occurrence predictions through which the system identified dangerous fire variations and location clusters. The synthesis work in the conclusion received help from Simin Bai.

Yongxin Ye performed studies on fire occurrences through time and places that frequently experienced fires within Boston. She utilized the Transformer-based model structure to characterize seasons while developing Time Series Analysis for future fire occurrence forecasting. Yongxin generated the Plotly Choropleth Map which displayed fire incident locations throughout various Boston neighborhoods. Fire occurrence patterns differed seasonally according to the analysis and demographic characteristics combined with infrastructure types contributed to fire risk across the city. Yongxin took charge of merging all sections in the project including Conclusion and Contributions as well as References.

Madhan Jothimani led an investigation to determine the key features influencing fire incident severity. Using Logistic Regression, the research identified factors contributing to substantial fire incidents, guiding effective resource strategy development. Madhan Jothimani was responsible for exploratory data analysis and the interpretation of all sections in the severity classification model, ensuring a comprehensive understanding of the data and the factors affecting fire incidents.

Team members jointly worked on creating the PowerPoint presentation to deliver results in an effective manner.

Reference

1. Baiardi, J., Braastad, A., & Ferland, K. (2023). Fire Department Reports in Boston, MA. University of Massachusetts Dartmouth. Retrieved from https://bpb-us-w2.wpmucdn.com/sites.umassd.edu/dist/4/1458/files/2023/12/522-Boston-Data.pdf
2. City of Boston. (2024, December 9). Fire Incident Reporting [Dataset]. City of Boston's Open Data Portal.https://data.boston.gov/dataset/fire-incident-reporting/resource/91a38b1f-8439-46df-ba47- a30c48845e06
3. Marco Casalaina. (2024, November 11). The Future of AI: Generative AI for Time Series Forecasting. Retrieved from https://techcommunity.microsoft.com/blog/aiplatformblog/the-future-of- ai-generative-ai-for-time-series-forecasting-a-look-at-nixtla-time/4289510
4. Milanović, S., Marković, N., Pamučar, D., Gigović, L., Kostić, P., & Milanović, S. D. (2021). Forest Fire Probability Mapping in Eastern Serbia: Logistic Regression versus Random Forest Method. Forests, 12(1), 5-. https://doi.org/10.3390/f12010005
5. Rathnayake, R. M. D. I. M., Sridarran, P., & Abeynayake, M. D. T. E. (2020, March). Factors contributing to building fire incidents: A review. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Dubai, United Arab Emirates (pp. 10-12). https://www.ieomsociety.org/ieom2020/papers/138.pd
6. IBM Technology. (2022). What is Random Forest? [Video]. YouTube. https://youtu.be/gkXX4h3qYm4