DATA 512 Project Final Report

Debbie Davis
4 Dec 2024

Introduction

Wildfire smoke is an increasingly pressing issue affecting even those communities that are far from the source of the fires. In Memphis, Tennessee, residents face indirect impacts from smoke exposure, which raises questions about potential disruptions to economic stability and labor market dynamics. This study directly addresses these concerns by exploring the link between wildfire smoke and unemployment rates, a real problem that could affect thousands of workers and businesses in the area.

The motivation for this analysis stems from the need to understand how environmental hazards—particularly smoke exposure—affect the workforce in urban areas like Memphis. With the frequency and severity of wildfires projected to increase, it is crucial for local policymakers to have access to data-driven insights. By studying the relationship between smoke exposure, weather conditions, and labor metrics, this research aims to help city officials proactively protect workforce availability and minimize negative economic impacts.

Background/Related Work

Several studies inspired this project's approach, providing context and methodological insights into analyzing the impact of wildfire smoke on labor market outcomes.
Borgschulte et al. (2022) explored the relationship between air pollution and labor market outcomes in their study, "Air Pollution and the Labor Market: Evidence from Wildfire Smoke." This work provided the foundational idea of using changes in labor market metrics as the dependent variable, mainly focusing on quarterly variations to capture the impact of environmental hazards over time. Evaluating how smoke exposure influences labor market conditions directly informed the decision to use unemployment rates as a key outcome metric in this study.

Nielsen-Pincus et al. (2013) examined the effects of large wildfires on employment, wage growth, and volatility in the western United States, using a GARCH model to understand the volatility in wage growth. While using a GARCH model was not feasible for this project due to time constraints, it nonetheless inspired several aspects of the study's design. Specifically, it highlighted the importance of incorporating environmental factors, such as weather data, into labor market analysis. This insight informed the inclusion of NOAA weather variables to provide a more comprehensive understanding of the factors that may influence employment dynamics during wildfire smoke exposure.

Meier et al. (2023) provided further context on the broader economic effects of wildfires in their study "The regional economic impact of wildfires: Evidence from Southern Europe." While this research did not directly influence the modeling approach used in the current study, it underscored the importance of understanding the economic impacts of wildfires on a regional scale. This informed the overall direction of this project, motivating an exploration of how smoke exposure could impact the Memphis labor market specifically.

In this study, two key research questions were explored to assess the broader implications of wildfire smoke on labor market outcomes:

1. **What is the impact of wildfire smoke exposure on unemployment rates in Memphis, Tennessee?** This question aims to quantify whether increased smoke exposure is directly associated with changes in unemployment, helping to understand the degree of influence environmental hazards have on labor market stability.
2. **How can environmental factors such as wildfire smoke and seasonal weather variables be used to predict labor market outcomes in Memphis?** This question focuses on the predictive aspect of the analysis, investigating whether a combination of smoke exposure, weather data, and economic metrics can effectively forecast changes in unemployment.

These research questions were guided by the findings and approaches of previous studies, informing the selection of variables, the use of weather data, and the modeling techniques applied in this analysis.

Methodology

This project involved building two distinct models: a Smoke Impact Model and an Economic Impact Model, both utilizing gradient-boosted decision trees. The primary objective was to estimate the future year-over-year percentage change in unemployment rates for Memphis based on historical smoke impact, weather, and economic data. The Smoke Impact Model was designed to predict smoke exposure levels, which were then used as features in the Economic Impact Model to assess their impact on unemployment rates.

Smoke Impact Model

The Smoke Impact Model was developed using USGS data on historical wildfires between 1984 and 2020 within a 650-mile radius of Memphis, Tennessee. This dataset included the following variables: fire size, distance, and the perimeter of each fire represented by latitude and longitude points. Please reference the link for the “Wildfire Dataset” in the Data Sources section at the end of this report. The distance from Memphis was determined by calculating the centroid of each fire, based on the provided list of latitude and longitude points, and measuring the distance from that centroid to Memphis. This simplified calculation allowed for a consistent method to determine proximity of all fires included in the analysis.

To quantify the contribution of each fire to local smoke exposure, a smoke impact score was calculated using two key variables: fire size (in acres) and distance from Memphis. Larger fires closer to Memphis were assumed to have a more significant impact on smoke exposure in the area. The smoke impact score for each fire is calculated as follows:

$$Smoke\ Impact\ per\ Fire = \frac{\text{Fire Size}}{\text{Maximum Fire Size}} \text{ x } \left(1 - \frac{\text{Distance from Memphis}}{\text{650 miles}}\right)$$

This formula accounts for both the size and distance of each fire, providing a standardized impact score. Fire size and distance are normalized to standardize the effects across different fires. The distance factor is calculated by subtracting the ratio of the distance from Memphis to the maximum considered distance from 1. This gives greater weight to closer fires, resulting in a larger smoke impact value.

The smoke impact was then amortized over the estimated duration of each fire, allowing for a more refined daily estimate of its impact. This amortization helps reflect the gradual release of smoke, as opposed to treating the impact cumulatively at a single point in time. The daily smoke impact is calculated as

$$Daily\ Smoke\ Impact\ per\ Fire = \frac{\text{Smoke Impact}}{\text{Estimated Fire Duration}}$$

This approach provided a more dynamic estimate of smoke impact that reflects real-time contributions and allows for effective comparisons across different fires, especially when considering multiple simultaneous events. Additionally, larger fires are given greater weight, consistent with the assumption that they produce more smoke, which travels further depending on proximity to Memphis.

After calculating the daily smoke impact for each fire, these impacts were aggregated on an annual basis. For each year, the total smoke impact was obtained by summing all daily impacts, while the total fire duration represented the cumulative number of days of all fires that year. The average daily smoke impact per year was then computed by dividing the total yearly smoke impact by the cumulative number of days of all fires that year, providing a standardized metric for comparison.For example, if the total smoke impact for a year was 100 units and the total fire duration was 50 days, the average daily smoke impact would be $\frac{100}{50} = 2\ units\ per\ day$. This metric helps illustrate the average contribution of fires to smoke exposure on a daily basis, specifically on the days when there were active fires.

The Smoke Impact Model was built using a gradient-boosted decision tree approach with the calculated average daily smoke impact score as the label and the following features: fire size and distance. One of the key reasons for using a decision tree model was its ability to handle complex interactions between features and to potentially extend beyond simple linear relationships. While the current score calculation directly involves fire size and distance, the model was designed with the flexibility to handle more complex relationships that could emerge if additional features were incorporated in the future.

Another reason for using this approach was the potential to adjust weights attributed to each feature, enabling the model to discover non-obvious relationships that the initial formula might not capture. Decision tree models, especially boosted ones, excel at automatically learning the most impactful features and their combinations, which could be particularly beneficial if different types of fires or environmental conditions needed to be factored into the analysis later on. Due to time constraints, the model in this iteration remained simple, but the framework is in place to explore these advanced options in future iterations.

To ensure the model's accuracy and reliability, hyperparameter tuning was conducted across multiple iterations, each with different combinations of settings, such as learning rate and maximum tree depth. The best-performing set of parameters was selected for the final model. Bootstrapping was then used to evaluate the stability of the model, with the model retrained 100 times on resampled subsets of the training data to assess the variability in R^2 scores. Future predictions of smoke impact from 2025 to 2051 were generated using 5-year rolling averages for the fire size and distance from Memphis features.

Economic Impact Model

The BLS data was obtained for each month from 2007 to 2024 for the Memphis, TN-MS-AR Metropolitan Statistical Area from the CES and LAUS datasets (please refer to the links in the Data Sources section). NOAA data was gathered for each month between September 2009 and September 2024 from the Global Summary of the Month (GSOM) dataset for the Memphis International Airport weather station (please refer to the link in the Data Sources section).

One challenge was that the BLS and NOAA datasets provided data at monthly intervals, while the predicted smoke impact scores from the Smoke Impact Model were only available on a yearly basis. To address this mismatch, the yearly smoke impact scores were converted into monthly values by repeating the yearly score for each month within the corresponding year. This approach allowed for a consistent number of data points across all variables, maximizing the available data for training the Economic Impact Model.

The Economic Impact Model used a gradient-boosted decision tree approach similar to the Smoke Impact Model. For training, the input features and the label (year-over-year percentage change in the unemployment rate) were filtered to the period from 2009 to 2020. This decision was based on the availability of historical data: wildfire data was only available through 2020, while NOAA data extended from 2009 to 2024.

To ensure reliability and optimal performance, hyperparameter tuning and bootstrapping were conducted. Hyperparameter tuning identified the best set of parameters for the model, which were used in the final version. Bootstrapping, involving 100 iterations with resampled data, was then applied to assess the variability of the R^2 scores and ensure robustness.

For future predictions, a slightly different approach was employed. Similar to the Smoke Impact Model, future values for the NOAA and BLS variables were estimated using a 5-year rolling average approach. However, for the smoke impact variable, the predictions for 2025 to 2051 were directly taken from the Smoke Impact Model itself. This integration allowed the Economic Impact Model to capture how future smoke exposure levels, in combination with economic trends, could affect unemployment rates in Memphis.

This methodology, by using the two models in tandem, provided a comprehensive approach to understanding the predicted smoke impacts and economic consequences for Memphis. Gradient-boosted decision trees allowed for the capture of complex patterns in the data, while hyperparameter tuning and bootstrapping ensured the robustness and reliability of the models.

Human-Centered Considerations

Ethical considerations and human-centered design were integral to this study. Data privacy was maintained by exclusively using publicly available datasets that contained no personally identifiable information. Additionally, potential biases in the data were considered carefully to ensure that the models did not inadvertently amplify existing inequalities. When selecting the Smoke Impact Model and the Economic Impact Model features, care was taken to avoid including features that could act as proxies for sensitive attributes, such as race, ethnicity, or socioeconomic status. For instance, focusing on environmental variables like smoke exposure, weather, and aggregated economic metrics (instead of neighborhood-specific data) helped reduce the risk of inadvertently encoding demographic disparities.

The study aimed to understand the relationships between smoke exposure and unemployment rates and provide policymakers with transparent, actionable insights. By clearly communicating the assumptions, limitations, and results of the models, the study aimed to support informed decision-making. If the model training or testing results, such as R^2 values, indicated that the model was unreliable or had poor predictive performance, this would be accurately conveyed to avoid any potential misinterpretation of the findings.

Changes from Part 2 - Extension Plan

Several changes were made during the project compared to the initial plans outlined in Part 2 - Extension Plan. Due to time constraints and limitations in data availability, adjustments were necessary to refine the scope of the analysis and focus on achievable outcomes within the project timeframe.

Firstly, while the original plan included analyzing both unemployment rates and total hours worked as labor market metrics, the final analysis focused solely on unemployment rates as the label for the Economic Impact Model. This decision was made to simplify the analysis and ensure it could be completed effectively with the available time and resources.

Another significant change involved the modeling approach. Initially, the plan was to use a linear regression model to determine the relationship between wildfire smoke exposure and labor market outcomes. However, the initial linear regression results were unsatisfactory, with negative R^2 scores indicating poor model performance. As a result, the approach was revised, and a gradient-boosted decision tree model was used for the Economic Impact Model. This change allowed for a better fit to the data and the ability to capture more complex patterns, ultimately yielding stronger performance metrics.

Data availability also impacted the planned analysis. Initially, the goal was to use the Current Employment Statistics (CES) data specific to the Trade, Transportation, and Utilities sector and potentially include other industry sectors in Memphis. Unfortunately, the CES data was only available at a broad level for the private industry, preventing sector-specific analysis. Consequently, the Economic Impact Model was limited to using the aggregated CES data for private industry, which restricted the ability to understand the sector-level impacts of wildfire smoke.

Lastly, the use of weather data also changed slightly from the initial plan. While the original extension plan intended to use multiple weather variables, including average temperature, maximum and minimum temperatures, and total monthly precipitation, only Average Monthly Temperature and Total Monthly Precipitation were ultimately included. This decision was based on the need to streamline the analysis and reduce the complexity of the model while maintaining sufficient environmental context to understand the impact of smoke exposure.

These changes balance the extension plan's initial ambitions with the practical limitations encountered during the implementation phase. Despite these adjustments, the project successfully developed insights into the relationship between wildfire smoke exposure and labor market outcomes in Memphis, providing valuable guidance for future policy decisions.

Findings

Smoke Impact Model

The predicted values for the input features (fire size and distance) were generated using 5-year rolling averages (reference Figure 1). These rolling averages were used to smooth out annual fluctuations and provide a more stable input for future predictions.

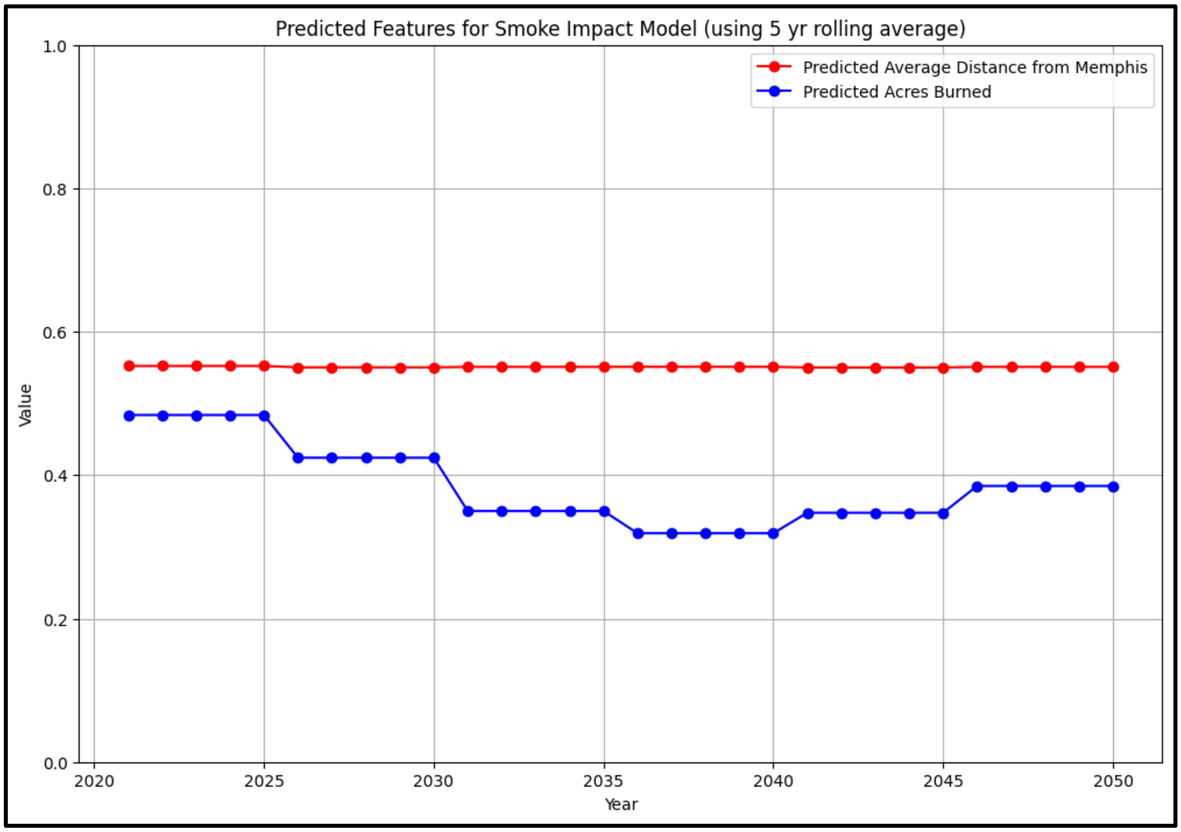

Figure 1: Predicted Values for Smoke Impact Model Features

The Smoke Impact Model showed considerable variability in its R^2 scores, which ranged from as low as 0.26 to as high as 0.97, with an average R^2 of 0.734 (reference Figure 2) . This spread indicates a lack of consistency across bootstrap samples, suggesting that the model may not generalize well. Given that the label for the Smoke Impact Model is directly derived from the features used to train it, this performance variability is not entirely surprising. Consequently, the

Smoke Impact Model predictions (reference Figure 3) should be interpreted with caution, as the model does not provide additional insight beyond the initial calculated values.

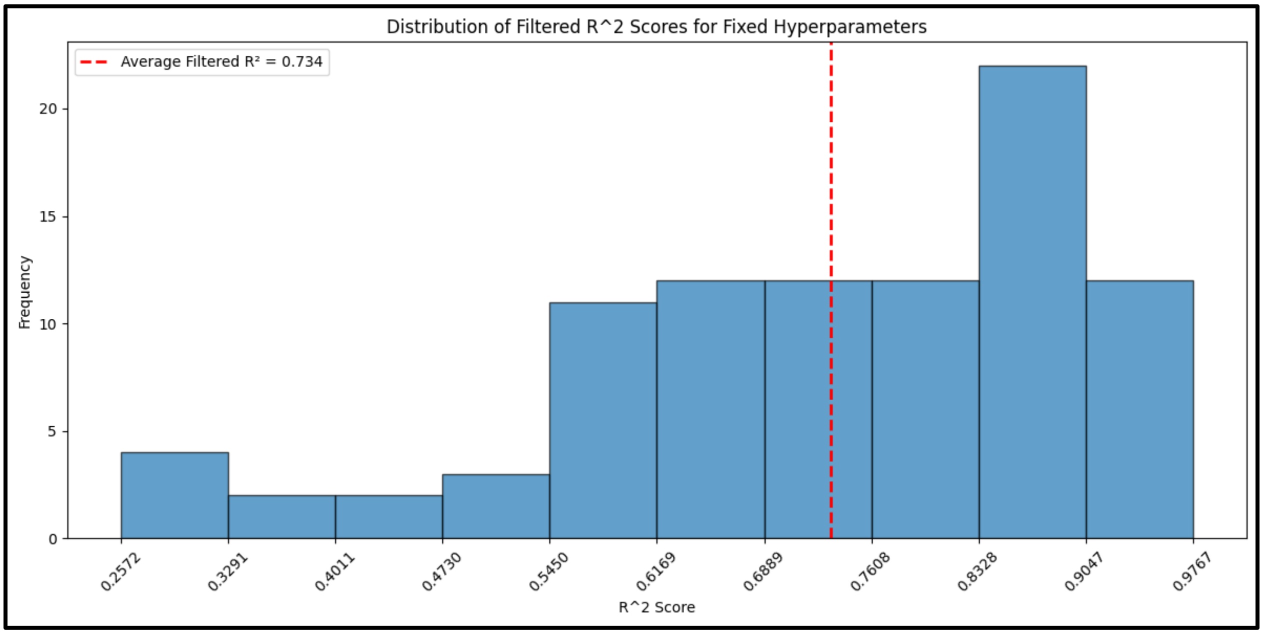

Figure 2: Bootstrap Iteration Results from Smoke Impact Model Training

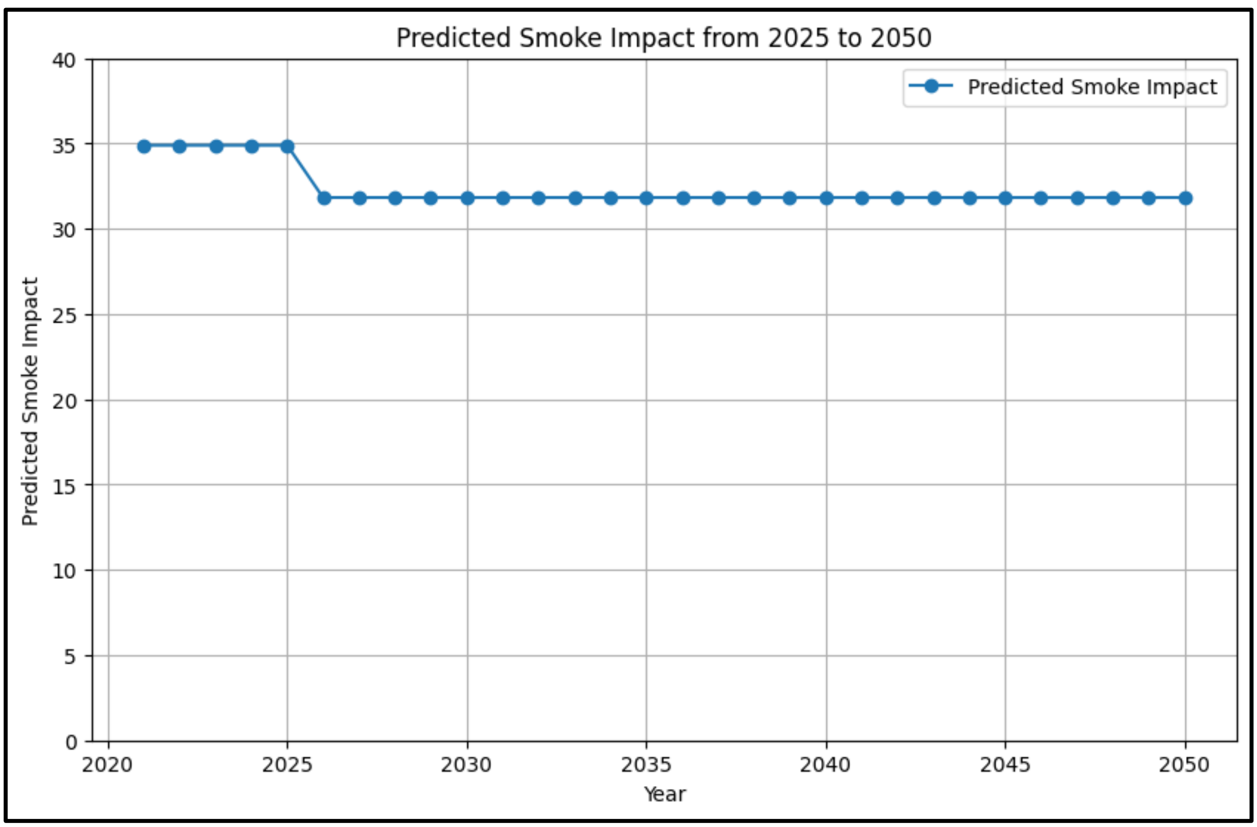

Figure 3: Smoke Impact Model Predictions

Economic Impact Model

The predicted values for the input features were generated using 5-year rolling averages (reference Figure 4 and 5). These rolling averages were used to smooth out annual fluctuations and provide a more stable input for future predictions.

The Economic Impact Model demonstrated strong and consistent performance. The R^2 scores were predominantly high, with most values above 0.8 and an average of 0.821 (reference Figure 6). This indicates that the model was able to reliably learn relationships between the input features and the target variable. The consistency of high R^2 scores suggests that the Economic Impact Model captures meaningful patterns and is effective in predicting economic impacts linked to smoke exposure. These findings indicate that the Economic Impact Model provides more reliable and actionable insights compared to the Smoke Impact Model. Economic Impact model predictions are shown in Figure 7.

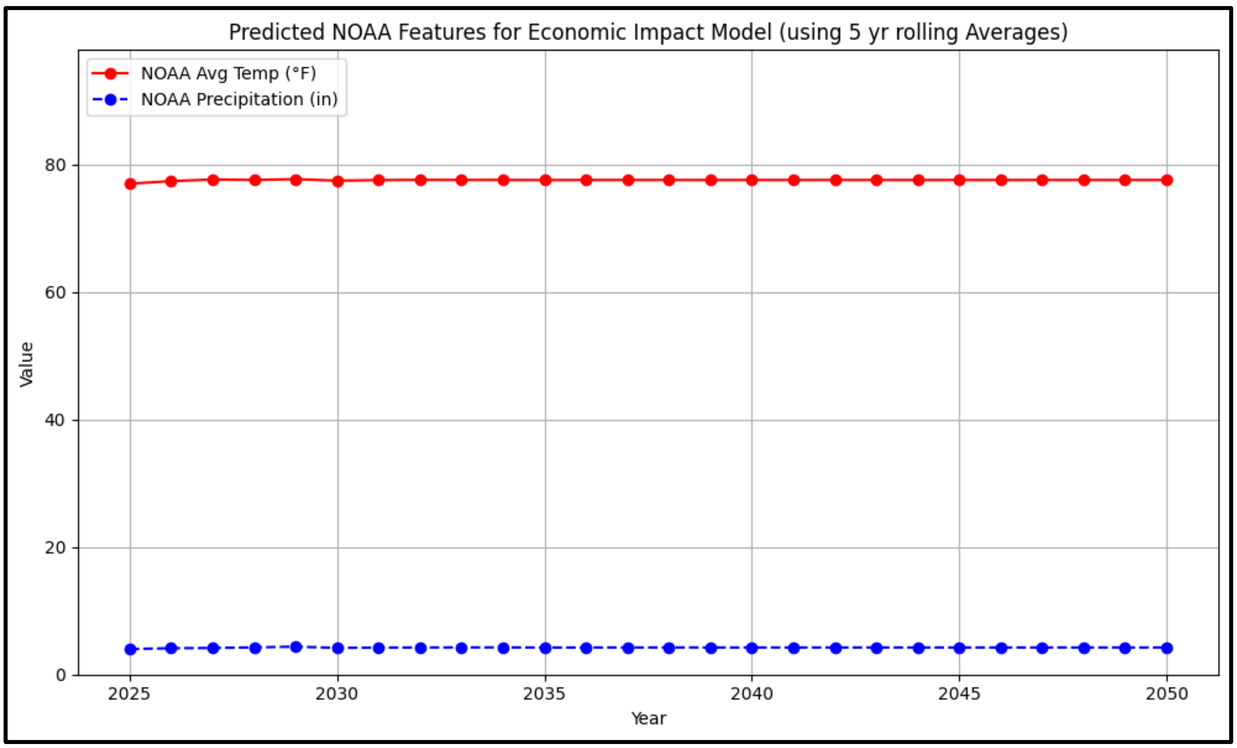

Figure 4: Predicted Values for NOAA Features for Economic Impact Model

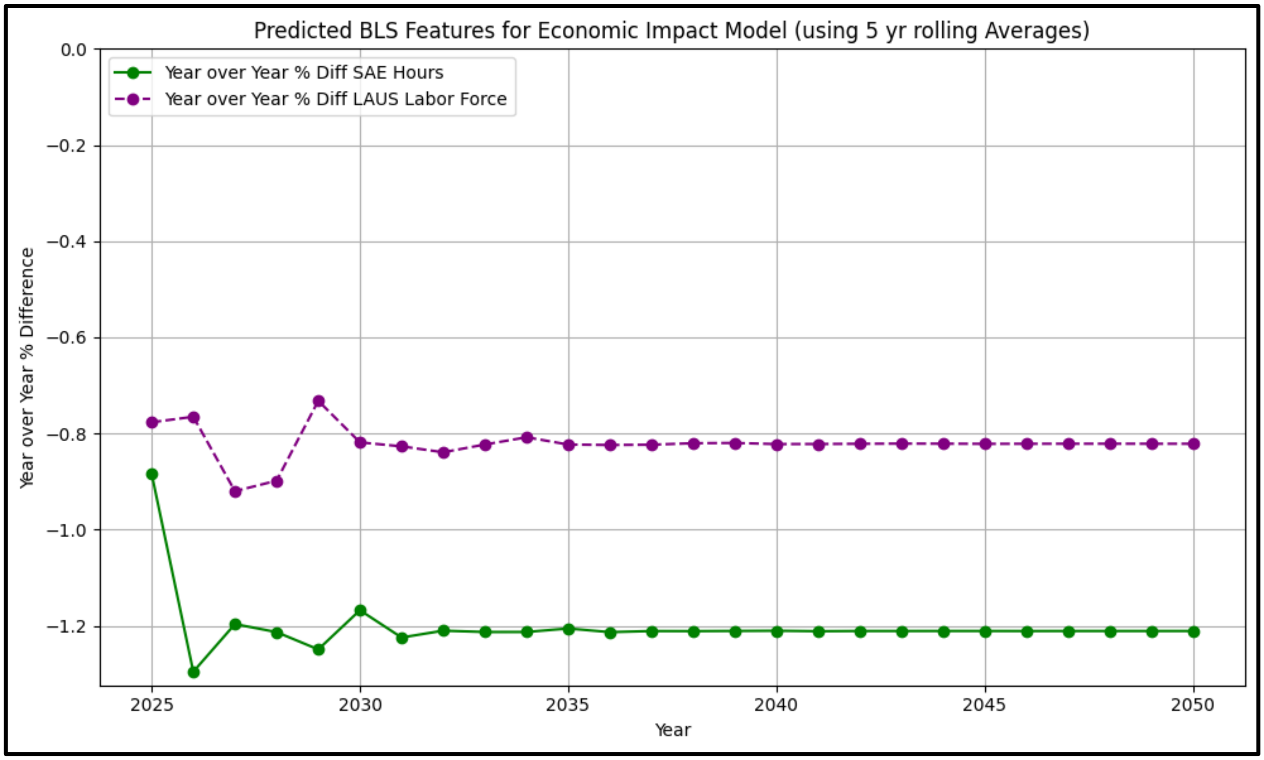

Figure 5: Predicted Values for BLS Features for Economic Impact Model

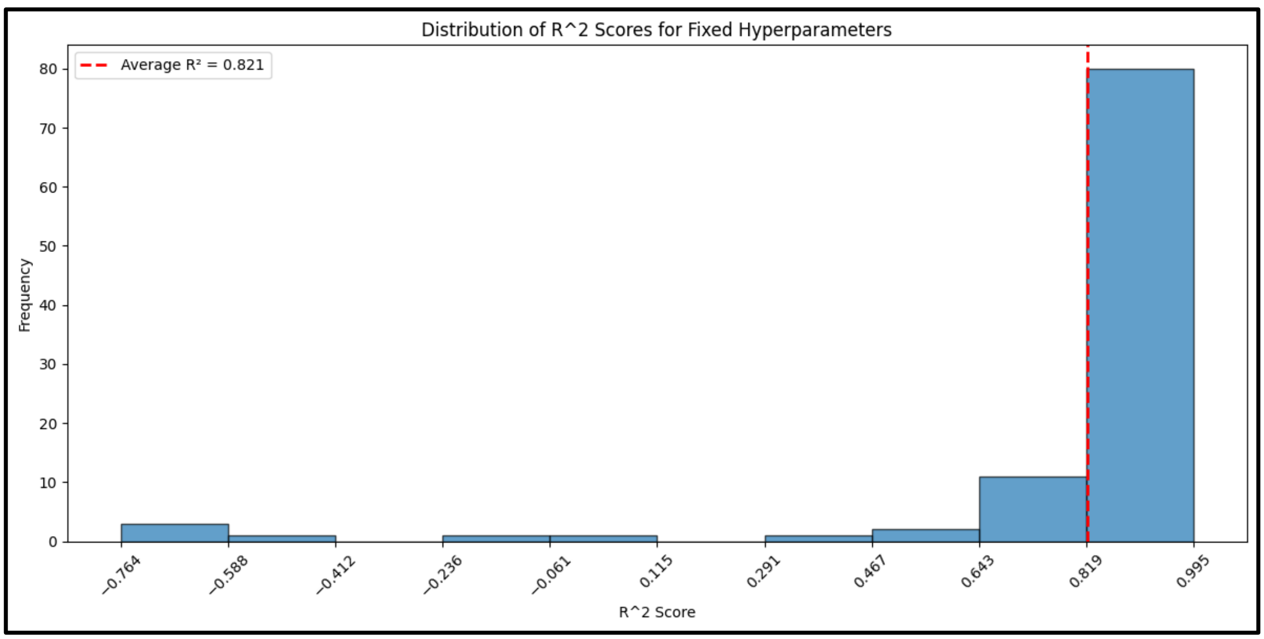

Figure 6: Bootstrap Iteration Results from Economic Impact Model Training

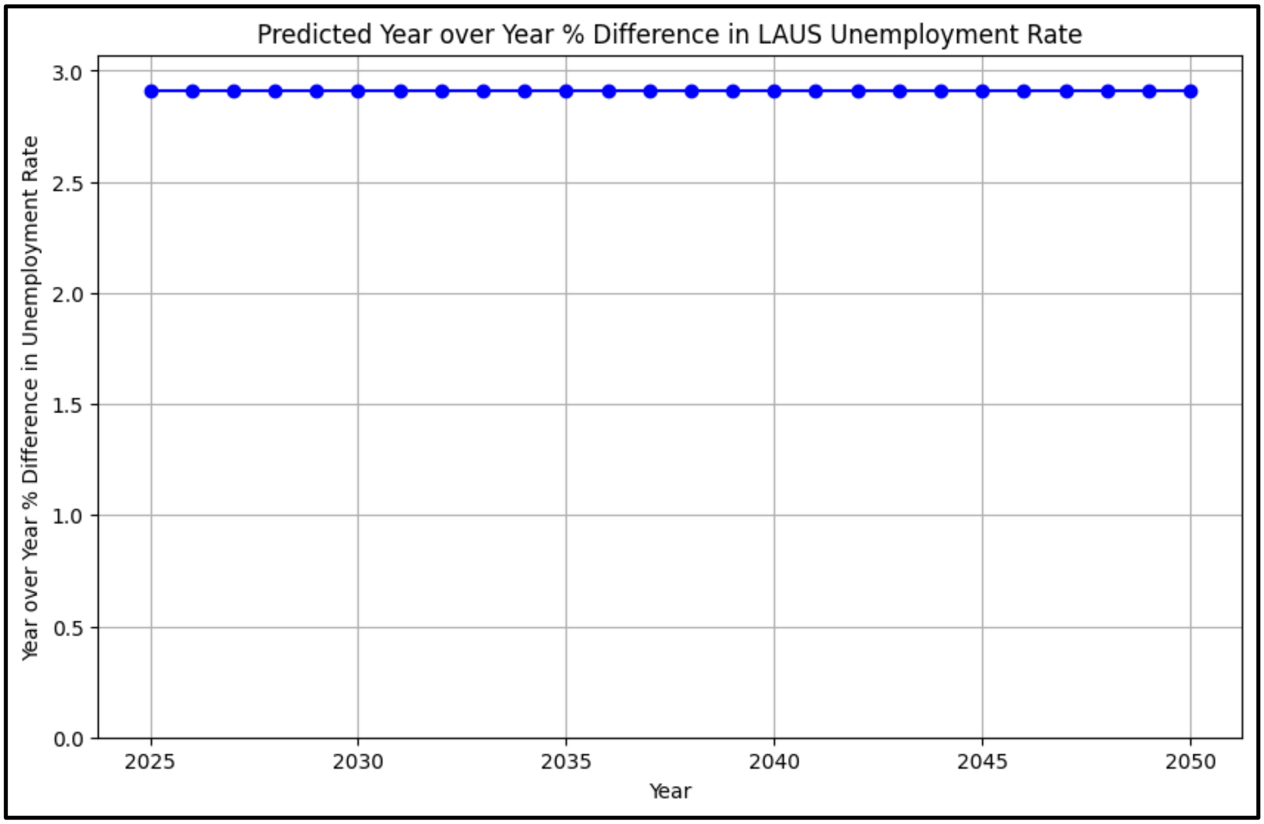

Figure 7: Economic Impact Model Predictions

Discussion/Implications

The consistent predictions from the Economic Impact Model, showing a steady 3% increase in unemployment over time, indicate that the impact of wildfire smoke may be subtle and driven by other underlying factors not included in the current model. The lack of variability suggests that additional features—such as broader economic metrics or health outcomes—might be required for a more comprehensive understanding of how smoke exposure impacts labor dynamics.

For city officials, immediate actions might include increasing community resilience by offering educational resources to both employers and workers on how to mitigate the effects of smoke exposure. Over the next year, policymakers should work on implementing air quality alert systems and encouraging adaptive measures such as reduced outdoor work during poor air quality days. Over a longer period (three to five years), the city could consider integrating more granular data on sector-specific employment metrics to refine future assessments and support industries more vulnerable to environmental impacts. Additionally, engaging with local communities to understand their needs and expectations will be critical in developing policies that not only protect economic stability but also support the well-being of Memphis residents

Human-Centered Considerations

Human-centered data science principles informed this project's various aspects, emphasizing transparency and equity. From the outset, transparency was prioritized by ensuring that the model's assumptions, limitations, and uncertainties were clearly communicated, notably when substantial variation in R^2 values indicated potential reliability issues. This allowed for honest interpretation of the results, ensuring stakeholders would not be misled by overconfidence in the model's outputs.

Equity was also a key guiding principle. Feature selection was conducted carefully to avoid variables that could act as proxies for sensitive attributes like race, ethnicity, or socioeconomic status. Instead, the analysis focused on environmental variables such as smoke exposure and general economic metrics, reducing the risk of unintentionally encoding demographic disparities. By prioritizing fairness in model development and providing insights that were both actionable and sensitive to the unique needs of Memphis' workforce, the project aimed to empower local decision-makers to create equitable interventions that mitigate the negative impacts of wildfire smoke, particularly for vulnerable populations

Limitations

This study has several limitations that impact the accuracy and reliability of the findings, primarily stemming from issues with the datasets used. First, the wildfire dataset presented significant limitations. Ideally, this dataset would include daily latitude and longitude points for the fire perimeter, allowing for a precise assessment of smoke impact using wind data to determine which areas were affected each day. Instead, the available data included only a single set of latitude and longitude points for the perimeter of each fire, which limited the precision of geographic analysis. Additionally, the fire start and end dates were unreliable, further complicating the alignment of wildfire events with other datasets. To simplify the analysis, the distance from Memphis was calculated based on the centroid of each fire, which may have resulted in underestimating smoke exposure for the city if fires impacted Memphis but had a distant center point.

Weather data from NOAA also presented challenges. Due to frequent outages on the NOAA website, I could only access data for 2009-2024, which limited the historical perspective needed for a comprehensive analysis. Data from 2007 and 2008 would have been beneficial for providing a baseline for pre- and post-recession trends, but their absence weakened the ability to capture long-term patterns.

Economic data, particularly from the Bureau of Labor Statistics (BLS), had its own limitations. Average Weekly Hours of All Employees were only available for private industry as a whole in Memphis, without the granularity to drill down into specific sectors like manufacturing. This lack of sector-specific data limited the ability to assess which industries were most impacted by wildfire smoke exposure. Additionally, the dataset included significant economic disruptions in 2008 and 2020. While their inclusion introduced variability, they were retained to maximize the dataset's size and provide insights into how economic crises interact with environmental stressors.

The lack of health data also limited the scope of this study. Due to the absence of publicly available health data for Memphis, health impacts from wildfire smoke could not be analyzed. Reaching out to the Tennessee Department of Health was considered, but it was impractical within the project's timeline. While using health data from other U.S. regions was an option, differences in population health and regional exposure could have led to misleading conclusions for Memphis.

Finally, gradient-boosted decision tree models, though effective at capturing complex relationships, are more challenging to interpret than simpler models like linear regression. This limitation limits the ability to provide straightforward, transparent explanations to stakeholders who may require clearer insights into the model's outcomes.

Conclusion

In conclusion, this study explored the relationship between wildfire smoke exposure and labor market outcomes in Memphis, using gradient-boosted decision tree models to predict smoke impacts and their potential effects on unemployment rates. The findings, which indicate a relatively stable but consistent increase in unemployment, provide initial insights into the subtle ways that smoke exposure may influence economic conditions.

This study was guided by human-centered data science principles, prioritizing transparency, community relevance, and ethical considerations throughout. The focus on clear communication of the models' strengths and limitations ensures that stakeholders—such as the Memphis town council—are equipped with information that respects the complexity of the problem without overstating predictive certainty. This approach helps avoid misleading conclusions that could result in inappropriate policy actions.

Future research should aim to incorporate a broader range of features, such as health outcomes and industry-specific data, to provide a more comprehensive view of how smoke impacts workforce stability. By emphasizing ethical transparency and striving for actionable insights that benefit the community, this research lays the foundation for developing proactive strategies that address both current and future risks to economic resilience in Memphis

References

1. Borgschulte, M., Molitor, D., & Zou, E. Y. (2022). Air Pollution and the Labor Market: Evidence from Wildfire Smoke. *The Review of Economics and Statistics*, 1–46. https://doi.org/10.1162/rest_a_01243
2. Meier, S., Elliott, R. J. R., & Strobl, E. (2023). The regional economic impact of wildfires: Evidence from Southern Europe. *Journal of Environmental Economics and Management*, 118, 102787. https://doi.org/10.1016/j.jeem.2023.102787
3. Nielsen-Pincus, M., Moseley, C., & Gebert, K. (2013). The Effects of Large Wildfires on Employment and Wage Growth and Volatility in the Western United States. *Journal of Forestry*, 111(6), 404–411. https://doi.org/10.5849/jof.13-012

Data Sources

Wildfire Dataset

BLS - LAUS Data

BLS - CES Data

NOAA GSOM