ML & Climate | Final Paper

Jennifer Oettinger (jmo2171), Aryaman Kejriwal (ak4395)

May 11, 2025

1 Abstract

Predicting wildfire growth is essential for disaster preparedness. Currently, most literature on wildfire risk assessment centers around computationally intense physics-based simulations or correlative models to predict regions with high risk of wildfire growth. After creating a dataset by combining US Wildfire data with meteorological data, we developed a set of causal models to predict wildfire size in the United States. First we searched for graphical causal models (GCMs) represented as Directed Acyclic Graphs (DAGs) to describe our causal models. Next we modeled our variables using linear regression, polynomial regression, XGBoost and MORD (Multi-class Classifier for Ordinal Regression) and a combination of these techniques. We evaluated the best structural causal model (SCM) with a random test-train split and then with a time-dependent test-train split to understand its generalization potential. We compared our SCM's performance on both types of test-train splits to the performance of an XGBoost model (our best performing traditional ML model) on the same task.

2 Introduction

The early twenty-first century has seen a marked escalation in the frequency and severity of wildfires [10]. In the United States, the annual number of wildfires rose from 140 in 1980 to 250 by 2012, with an average of 1.11 billion acres burned each year [2, 1]. Wildfires have produced a sharp increase in human and animal casualties, property damage, and long-term ecological degradation [13].

The "fire triangle", which comprises a heat source, fuel, and oxygen are the elements needed for a wildfire to break out. Natural phenomena such as lightning can deliver sufficient heat to ignite dry vegetation. The availability of combustible fuel ensures that small ignitions can escalate into widespread fires [8]. Wind plays a critical role in transporting embers and intensifying fire spread, particularly under turbulent or erratic wind patterns [14]. The growing prevalence of wildfires aligns closely with observed shifts in global environmental conditions. Rising global temperatures, prolonged heatwaves, and persistent droughts, exacerbated by anthropogenic climate change, are key contributors to the increase in wildfire activity [11]. Higher temperatures extract moisture from soils and vegetation, making landscapes more flammable. Changing atmospheric dynamics contribute to more volatile wind behavior, heightening the risk of fire expansion. These features have complex relationships with each other making the prediction task challenging.

To mitigate the devastating impacts of wildfires, it is essential to develop predictive tools capable of modeling fire growth. Current Machine Learning models used to predict wildfire size have fair performance. Recent work shows that Logistic Regression can attain an f1 score of 0.55 whereas Random Forest approaches can attain f1 scores of 0.56 [15]. Some dense data driven approaches such as Deep Learning, Decision Trees and Extreme Gradient Boosted Trees can even obtain accuracy scores of around 80% [5]. Despite these promising results currently used ML models are, for the most part, purely correlational black box functions. This makes it challenging to justify actions in the real world taken as preventative measures or planning of disaster response.

The goal of this paper was to create causal models which are able to model the relationships of predictive features to identify key environmental and meteorological drivers of fire growth. Such insights are vital to design effective interventions aimed at reducing the incidence and destructiveness of future wildfires.

3 Dataset

Initially we planned to predict the next day spread of a fire based on the current day fire location. To accomplish this task we planed to use the Next Day Wildfire Spread Dataset [9], which consists of two consecutive binary maps of fire location and 11 geographic and weather related geospatial features in a 64x64 km region. However a widely accepted framework for fitting causal models to data in this form could not be found, so we pivoted our project to a prediction task that was not geospatial - predicting ultimate size class from early information about fires.

Because no tabular dataset that combines geographic and weather data with fire occurrences and burn area could be found, a custom dataset was created for this project. We procured data on forest fires in the United States between 2000 and 2015 from the Fire Program Analysis (FPA) Fire-Occurrence Database (FOD) [12] maintained by US Forest Services. This database provides information such as the **date of fire start, date of fire containment**, **latitude**, **longitude** and **fire size classification**. As our project focused on predicting ultimate fire size from initial data, fires that were contained within one day offered little predictive value. We filtered out such cases and were left with about 60,000 fire incidences. As an additional bonus, the filtration step helped reduce class imbalances in the dataset.

The size of a forest fire is measured in the total number of acres of land burned by the fire. In addition to the number of acres the FPA FOD lists a size classification for each fire, based on the land area burnt using the classification scale in Table 1. This is the standard scale used for fire classification in the US. This classification will be used as the predicted feature for all our models.

Class	Area Burned (acres)
A	≤ 0.25
B	0.29 – 9.90
C	10.0 – 99.9
D	100 – 299
E	300 – 999
F	1000 – 4999
G	≥ 5000

Table 1: Wildfire classification scale based on land area burned (U.S. system).

We augmented our selected features from the FOD with predictive weather and geographic factors using the Google Earth Engine, many of which are inspired from those included in our original Next Day Wildfire Spread Dataset. From the GRIDMET DROUGHT: CONUS Drought Indices dataset [4] we obtained **Drought Index (PDSI)**, from the GRIDMET: University of Idaho Gridded Surface Meteorological Dataset [4] we obtained **Humidity**, **Precipitation**, **Wind Speed**, **Max Temperature**, **Min Temperature**, **100 and 1000 hr Dead Fuel Moisture** and **Energy Release Component**, from the MOD13A2.061 Terra Vegetation Indices 16-Day Global 1km dataset [7] we obtained **Vegetation (EVI)**, and from GPWv411: Population Density Dataset [6] we obtained **Population Density**. Each feature in our dataset is the average of the samples at a 10km scale around the initial fire location on the first reported day of fire (or most recent day before the fire in the dataset).

Our models take as inputs the geographic and weather data and predict the final size class of the fire. We predicted the size class instead of the exact size of the wildfire to eliminate unnecessary noise, such as signals from variables not included in our model, measurement errors and inconsistencies in determining what pieces of land are burnt and inconsistencies in the resolution at which the burnt land area is measured. Additionally knowing the size class

of the fire instead of exact burn area is sufficient to understand the causal relationship between variables and to plan useful preventative or disaster response actions.

3.1 Feature Analysis

After the dataset was prepared, feature analysis was conducted. A correlation matrix was fitted among all features (Figure 1). The only highly correlated features are minimum and maximum temperature and 1000 Hour Dead Fuel Moisture and Energy Release component (derived measures). For this reason many of our models use an average of minimum and maximum temperature, and don't use both derived measures. The distribution of features was also analyzed. The primary challenge was a severe class imbalance among fire size classes, with many more small fires than large ones.

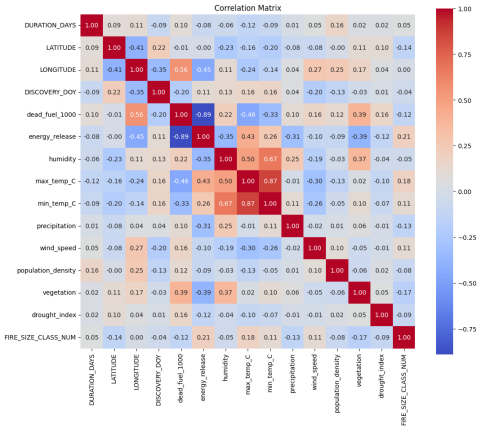


Figure 1: Correlation Matrix of Extracted Features in Dataset

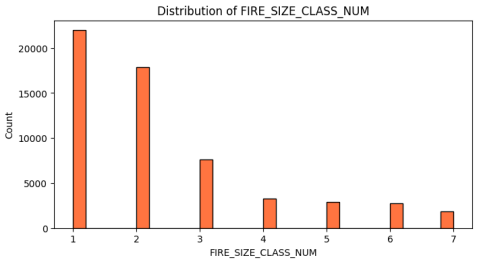


Figure 2: Distribution of Fire Classes (Class A=1, ..., Class G=7)

4 Methodology

4.1 Standard Machine Learning Models

To establish a non-causal baseline for our prediction problem, we used the following traditional machine learning (ML) models, (1) Feed Forward Neural Network, (2) Support Vector Machines (SVM), (3) Extreme Gradient Boosting (XGBoost), and (4) Ordinal Light Gradient Boosting (Ordinal LGBM). For each method, class weighting was used in the training of the models in order to attempt to account for the size class imbalance.

4.1.1 Feed Forward Neural Network

A Feed Forward Neural Network is a type of artificial neural network in which information flows uni-directionally from input to output through hidden layers. Each layer consists of interconnected neurons that apply learned transformations to the input data.

4.1.2 Support Vector Machine (SVM)

A Support Vector Machine is a supervised learning algorithm that constructs decision boundaries to separate data points belonging to different classes. The algorithm maximizes the margin between the classes, often leading to better generalization.

4.1.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is an optimized implementation of gradient boosted decision trees. It builds an ensemble of decision trees in a sequential manner, where each subsequent tree attempts to correct the errors of its predecessors.

4.1.4 Ordinal Light Gradient Boosting (Ordinal LGBM)

Ordinal Light Gradient Boosting is an adaptation of the Light Gradient Boosting Machine (LightGBM) framework designed specifically for ordinal classification tasks. LightGBM is similar to XGBoost in that it constructs an ensemble of decision trees to make predictions. Unlike the other traditional ML methods mentioned above, which make purely categorical predictions, Ordinal LGBM incorporates the ordered nature of the target by optimizing loss functions that respect the inherent ranking between classes, capturing the class identity and the relative order among classes.

4.2 Causal Models

Causal models are used to derive causal relationships from observational data. We proposed causal structures represented as directed acyclic graphs (DAGs). We posited candidate graphs and used causal learning algorithms to produce a set of DAGs. Conditional independencies in our data were used to reject DAGs which violate these casual observations in the data. After identifying the best DAG, we tested the performance of linear models, polynomial models, MORD and XGBoost Models and a mix of these approaches in predicting the relationship between the variables in the DAG.

4.3 Generalization Study

The increase in the size and frequency of wildfires alludes to a distribution shift under which models trained on historical data may not adequately model future fires. To asses how both our best traditional ML model and best causal model perform under this distribution shift, we conducted a generalization study. Instead of splitting our data in training and test sets randomly, we split it based on year. All fires that occurred prior to 2014 were used in the training phase and so testing data consisted only of fires occurring in 2014 or 2015. This method of time splitting resulted in roughly the same proportion of data allocated to training vs testing (80% vs 20%) as in our random split validation.

4.4 Evaluation Metric

For the task of wildfire size prediction, recall and precision are the most important metrics. A high recall implies that there are many true positives and few false negatives. A high precision implies that there are many more true positives than false positives. We used F1-scores, a commonly used metric that combines recall and precision, for all model evaluations.

5 Results and Analysis

5.1 Traditional ML Model Determination

The F1-Scores for all four types of trained models can be seen in Table 2. XGBoost outperformed the other models, therefore we will use XGBoost as our traditional ML model baseline for the remainder of this report.

Model	F1-Score
Feedforward Neural Network	0.21
Support Vector Machine	0.24
XGBoost	0.28
Ordinal LightGBM	0.21

Table 2: F1-scores of evaluated models on the classification task.

The F1-scores for all models are relatively low, especially compared to other work [5, 15] which achieved f1-scores around 0.56. However, these results were obtained with a different set of

predictive features, such as US state and cause of fire ignition, which we did not see a meaningful way to analyze causally.

5.2 Search for Causal Relationships

The first step in our model development process was determining the best DAG to represent the causal relationships between our models. We began by using causal graph learning algorithms implemented in python's causal-learn library. This approach created massive nonsensical graphs that violated a lot of the causal assumptions. We then transitioned to using manually posited graphs using domain information. After conducting a literature review, we were not able to find conclusive research on the causal effects of all variables on each other and on wildfire spread in the US, since climate variables tend to have complicated non-linear effects on each other. They generally include interaction effects as well. However we were able to find small subsets of clear interactions between variables.

From these variables we tried different causality graphs that made sense intuitively. Initially, we found graphs with extremely poor predictive power. After iteratively searching through potential graphs, we were able to introduce 7 predictors into our DAG without violating causal assumptions. These predictors were **Humidity**, **Precipitation**, **Wind Speed**, **Average Temperature**, **Population Density**, **Drought Index**, and **Vegetation (EVI)**. We found the DAG in Figure 3 that did not violate our data's causality assumptions and made sense with the domain information we were able to collect about weather and geographic features. We used this DAG for the remaining causal analysis.

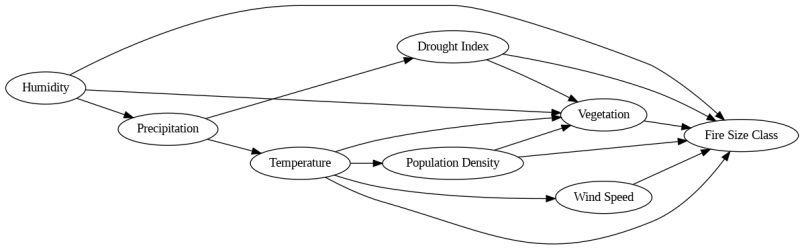


Figure 3: Final Causal Graph

5.3 Causal Function Determination

The next step was finding the right prediction function to model the relationship between the variables in our DAG.

5.3.1 Linear Model

We first constructed our SCM by assuming a linear relationship between all our variables. The benefit of a linear model is its high degree of interpretability. However, this model is limited in its degree of expressiveness. Wildfire size usually has a complicated relationship with its variables including interaction terms and nonlinearities, which as the name suggests can't be captured by linear models. Our linear model had f1-score $= 0.185$.

5.3.2 Other Polynomial Models

We further expanded on our model by allowing higher order polynomial terms in the variable relationships. We conducted quadratic, cubic and quartic regressions. Surprisingly, our model performance decreased marginally with the introduction of higher order terms. Our quadratic model had f1-score $= 0.183$, quadratic model had f1-score $= 0.182$ and quartic model had f1-score $= 0.183$. Since these models did not perform better than our linear model and are less interpretable, we decided not to use them.

5.3.3 MORD and XGBoost Model

We decided to take a different approach by using MORD to predict Fire Class Size to account for the fact that the variable is ordinal: there is an order to the discrete categories, however the difference between categories is not constant. We used XGBoost predictors to represent the relationships between other variables as XGBoost can take interaction effects into account. This model had an f1-score $= 0.184$ and did not outperform the linear model. Since it is also less interpretable than the linear model, we decided not to use it.

5.3.4 Mixed Model

While the XGBoost and MORD model didn't perform well to predict Fire Size Class, it was successful at predicting the Temperature, Wind Speed and Precipitation. So, we created a mixed model which would use XGBoost to predict these variables but a linear model to predict Population Density, Vegetation and Fire Size Class.

While we did see higher performance for the other variable predictions, Fire Size Class prediction had f1-score $= 0.184$. Once again, due to a lower f1-score and reduced interpretability, we did not use the Mixed model.

5.3.5 Other Approaches

Some other approaches we tried included predicting the actual fire size instead of the class, predicting the log of the fire size as well as using Ridge Regressions. None of these methods produced valuable outputs.

5.3.6 Final Model Specifications

Given the results of the aforementioned experiments, we chose to use a linear model to create out SCM. While the f1-score of the linear model is relatively low, it is extremely interpretable.

5.4 Model Evaluation

5.4.1 Random Split Evaluation

We first evaluated our causal model using a random test train split. Our causal model was compared to our best performing traditional ML model (XGBoost) which had an f1-score of 0.28. In this scenerio, our causal model had a testing data f1-score of 0.184. While our baseline correlational model performed substantially better than causal model, our causal model provides more explainable insights and informs us of how potential interventions may affect the final wildfire growth.

Our results are a proof of concept, demonstrating that more complex versions of our model may inform interventions that can be made in the short term or over the long term. An example of a short term intervention is removing vegetation from areas around the forest fire or high likelihood fire locations to limit growth. In the long run, we could intervene by enacting policy to incentivize reducing population density in fire prone regions or enacting policy that affects local and/or global temperature.

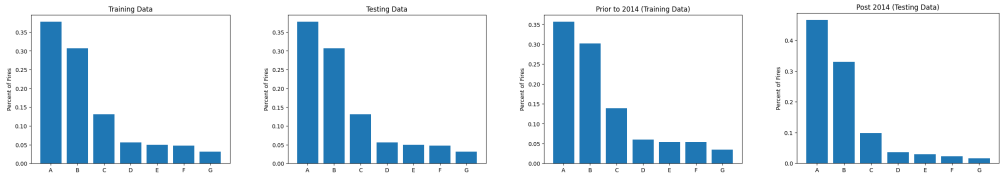
5.4.2 Time Generalization Study

Aligning with our methodology to understand the generalization of our results, we split the data such that our training split contained data from years prior to 2014, whereas the test split contained data from 2014 and beyond. This resulted in a split of about 80% of our data in our training split and about 20% of our data in our testing split, which resembles the proportions from the random split.

Similar to the random Test-Train split, our causal model performed worse than the traditional correlational model on the prediction task. However, notably, performance of the causal model on the time-dependent split increased relative to the random split, while the traditional ML

model's did not. Our causal model exhibited f1-score $= 0.225$ and the tradition ML baseline had an f1-score $= 0.27$. This result was counterintuitive, as we expected that performance on predicting out of distribution for both models would be worse.

Upon further exploration, we realized that while the fires per year of all categories was higher in the test split as compared to the training split, the number of class A and B fires was significantly higher. This resulted in the proportion of class A and B fires in the test split being even higher than that in the training split. One of the problems with our dataset was the class imbalance caused by the high proportion of smaller fires. So, our class imbalance is worse in our test split than it is in our training split, or the random split.



(a) Training (Random) **(b)** Testing (Random) **(c)** Training (pre 2014) **(d)** Testing (post 2014)

Figure 4: Class imbalance among random and time dependent train test splits

The higher class imbalance can explain our time-dependent test-train split results for causal models. Since our training split is biased towards predicting fires of classes A and B (since they are overrepresented in our dataset), the test split containing more of classes A and B means the causal model gets the prediction right a higher percentage of the time. Since XGBoost specifically accounts for the class imbalance, the apparent performance boost due to class imbalance wasn't seen to the same degree.

6 Discussion and Conclusion

While the traditional ML models performed better than the causal models, they still have relatively low f1-scores. To improve models on this prediction task as a whole, the collected dataset should be improved. The National Wildfire Coordinating Group [3] found that the critical causes of fires and particularly large fires is not just dry or hot environments, but rather *unusually* extreme conditions for the area. To improve our dataset and capture this relativity, in addition to the weather data at a snapshot in time when the fire first began, supplementing each observation with average conditions for a given location and day of year over a longer time period may be able to give better insight into the fires. Because the dataset covers such a large geographic span (the entire United States) with many different climates, adding historical averages for values may help models adapt to this large distribution of locations rather than only using absolute features which would vary greatly between areas of the country.

When fitting our causal models, the best performing models assumed a linear relationship between the variables. Due to the complex nature of the interaction between these variables and the rate of wildfire spread, more complex relationships between the variables than the ones studied in this paper should be analyzed to find models with better predictive power.

Future work may also include incorporating the predictions made by general climate models. Model predictions for variables like future temperature and precipitation may be incorporated as interventions in the wildfire model to understand how wildfire sizes are expected to change in the future. The conceptual limitation of our project in this application is that we only predict wildfire size given an initial wildfire, with no information regarding the probability of a wildfire in a region if no fire has started. While this might seem like a subtle change, it means that our model cannot be used to evaluate the frequency of forest fires, only the scale of a fire. A possible extension of our work could be predicting the probability and size of forest fires in a region over a year. Such a model in combination with the outputs of a powerful climate model will be helpful in predicting how the frequency of forest fires is expected to change over time.

Ultimately our our causal model does not perform as well as the current state of the art traditional ML models for predicting forest fire size. Traditional correlation based models are likely to consistently outperform causal models at the fire size prediction task and so should always be used for short term risk assessment. However, we found a GCM with good causal fit to our dataset that is consistent with literature we reviewed about the causal effects of climate variables on each other. We believe this GCM can be used as a stepping stone to create causal models with better predictive power, which can be used to identify long term forest fire risks and inform intervention design to reduce the risk of large wildfires.

References

[1] Facts + Statistics: Wildfires | III.

[2] Western Wildfires and Climate Change | Union of Concerned Scientists.

[3] Weather: Critical Fire Weather, April 2024.

[4] John T. Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3413.

[5] Mohammad Khaled Al-Bashiti and M. Z. Naser. Machine learning for wildfire classification: Exploring blackbox, eXplainable, symbolic, and SMOTE methods. *Natural Hazards Research*, 2(3):154–165, September 2022.

[6] Center For International Earth Science Information Network-CIESIN-Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11, 2017.

[7] Kamel Didan. MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V061, 2021.

[8] Kuo-Ming Hung, , and Ting-Wen Chen. A Novel Hierarchical Wildfire Alarm System Based on Vegetation Features,ERICDATA.

[9] Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next Day Wildfire Spread: A Machine Learning Data Set to Predict Wildfire Spreading from Remote-Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. arXiv:2112.02447 [cs].

[10] J. Rubí and Paulo R. L. Gondim. A performance comparison of machine learning models for wildfire occurrence risk prediction in the Brazilian Federal District region. *Environment Systems and Decisions*, 44, August 2023.

[11] Jennifer Sherry, Timothy Neale, Tara K. McGee, and Maria Sharpe. Rethinking the maps: A case study of knowledge incorporation in Canadian wildfire risk management and planning. *Journal of Environmental Management*, 234:494–502, March 2019.

[12] Karen Short. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_fod_20170508] (4th Edition).

[13] Catherine Stephenson, Handmer , John, , and Robyn Betts. Estimating the economic, social and environmental impacts of wildfires in Australia. *Environmental Hazards*, 12(2):93–111, June 2013. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17477891.2012.703490.

[14] Ali Tohidi and Nigel B. Kaye. Stochastic modeling of firebrand shower scenarios. *Fire Safety Journal*, 91:91–102, July 2017.

[15] Sky B. T. Williams. Wildfire Destruction — A Random Forest Classification of Forest Fires, July 2022.