Boston University Questrom School of Business

NOAA NIDIS Capstone Project:

# Predicting Wildfire Severity and Its Economic Impact on New England Communities

*Team B8: Ting-Hsuan Chen, Risheng Guo, Kratik Gupta, Jiazhi Jia, Maria Stella Vardanega*

# Table of Contents

# Report Summary

This project focused on the New England region as it has seen an overall increase in droughts and wildfires in recent years. After plotting the frequency of wildfires and the average DSCI per year over the past 20 years we noticed that on average there have been more wildfires and drought in the past 10 years compared to the previous 10. Furthermore, because the risk has yet to be fully realized by residents this report may be used as a mitigation approach for wildfire-prone communities in the New England region.

The first goal of this project was to predict wildfire severity in each county through machine learning models that could predict both fire size and fire class. After adding the average DSCI score of the previous 4, 8, and 12 weeks to the wildfire dataset as predictors, we used Random Forests for both regression and classification as we gained the best results in terms of prediction accuracy. For the classification model, our base accuracy was approximately 17% as there are 6 labels and thus anything less than 17% accuracy would mean the model was guessing. Our final model was able to predict the fire class with approximately 62% accuracy. Despite the positive results for the classification model, we discovered that the regression model had a negative r squared which illustrates that it performed worse than just predicting the mean fire size. This result was expected as the severity of a wildfire is caused by fuel, weather, and topography (Moore, 2021). Our model only took into account part of the weather attribute, meaning that more predictors are needed to successfully and accurately predict fire size.

The second goal of this project was to shed light on how wildfire costs impact various communities through a comparison of cost versus median income. We estimated the cost of a wildfire according to a household using insurance estimates. We observed the impact of wildfire cost on an average household by calculating the proportion that estimated wildfire cost represented in median household income. We found that in New London County, Connecticut where wildfire costs made up approximately 30% of the median income, the poverty rate was approximately 8% and 75% of the county encompassed white residents. Conversely, when looking at the county where the proportion was highest, Lamoille County, Vermont where the wildfire cost made up 67% of the median household income, we can observe some differences. We found that the poverty rate was similar whereas the white population was approximately 20% higher and the Hispanic/Latino, Black/African American, and Asian communities encompassed 10%, 6%, and 3% less of the residents respectively compared to New London County. This illustrates that for those counties in which households would struggle more heavily with recovering financially after a wildfire there are demographic disparities.

This project has applications in damage limitations, with FEMA, and in the insurance sector. In the case of damage limitations, our models can serve as the basis to predict the fire size which can aid communities in knowing whether to expect a smaller or larger wildfire which would help the preparation for damage limitations for each county. Similarly, for FEMA, which is working towards a fire-safe America, this project would shed light on differences between communities that may impact what safety measures and aid communities need. Last, our findings can apply to the insurance sector in calculating appropriate premiums.

Our project has potential for future growth in that with more explanatory variables in the dataset more accurate predictions can be achieved and be used to help counties mitigate fire damage by being prepared. Furthermore, by looking at the disparities between counties and the differing effects of wildfires US Fire Departments and county municipalities can take steps in ensuring that counties are equitably prepared for wildfires and the effects mitigated in a fair manner. For example, more aid could be provided for those in counties that are more prone to wildfires and particularly underserved in terms of income.

*Section references present in the text refer to sections in the code notebook which can be found on GitHub

# Introduction

## Background

Issues that have once been more prevalent in the Western United States, like wildfires, are quickly becoming a problem in New England as temperatures are rising and droughts are increasing with global warming and climate change. However, the more moist climate of the New England region has historically made intense wildfires a less common occurrence than on the West Coast. Residents in the New England region thus do not perceive the risk of wildfires. On top of this, because of the socioeconomic barriers, wildfire vulnerability is spread unequally across races and ethnicity. Communities that are mostly black, Hispanic, or Native American experience 50 percent greater vulnerability to wildfires compared with other communities. This is mainly because their neighborhoods are located in low-lying, less-protected areas, and many people lack the resources to evacuate safely. Moreover, their lower-income level makes it especially hard for these communities to recover after a large wildfire (Davies et al., 2018).

## Goals

The first goal of this project aims to provide a wildfire severity prediction system that illustrates the severity of wildfires in New England regions as a function of the drought conditions in the previous months.

Our second goal is to shed light on the difference in the impact that wildfires have on underserved communities when compared to other communities by illustrating the percentage of income that wildfire costs take up for specific counties in the New England region of the United States.

Finally, we hope to spread more awareness of the disparities between these communities that are prone to wildfires in hopes that municipalities can allocate resources accordingly.

# Methodology

## Collecting the Data

We first collected the wildfire dataset from the U.S Department of agriculture's forest service research archive. We downloaded and stored the wildfire data via SQLite as a CSV file. The wildfire dataset contains wildfire entries throughout the country, dating from 1992 to 2018.

Drought data for each county in New England was then obtained from the U.S. Drought Monitor website as a CSV file. The earliest data available was from 2000; therefore, we took drought data for each New England county from 01/01/2000 to 01/01/2020. The drought dataset contains the Drought Severity and Coverage Index(DSCI), which will be used to determine the drought conditions.

After finding these datasets, we looked through census data that would allow us to gain insight into the household income and the demographics of each county as well as the cost of wildfire damage. The census dataset was found through the census.gov website and stored as a CSV file.

# Exploratory Data Analysis

## Wildfire dataset

### Nationwide

After exploring the wildfire dataset, we found that from 2000 to 2018, the five states with the most wildfires cases are California(CA), New York(NY), Georgia(GA), Florida(FL), and North Carolina(NC). The five states with the fewest wildfires cases are Hawaii(HI), New Hampshire(NH), Delaware(DE), Vermont(VT), and Iowa(IA) (Figure 1).

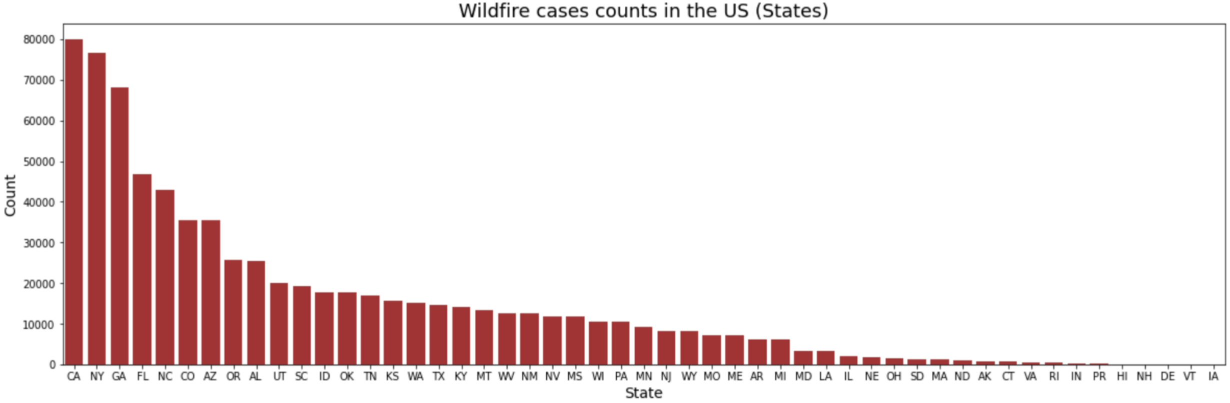

*Figure 1: Wildfire Cases in United States*

We found that wildfires mostly occurred in autumn and rarely occurred in winter. Next, we count wildfire cases of a certain fire class between 2000 to 2018. We found that fire incidents across the nation are increasing year by year, proving that this issue is worth noting (Figure 2). Since Maine, Massachusetts, and Connecticut are the NewEngland states with the most wildfires we focused on those states for our more specific analysis (Section 4.1.1).

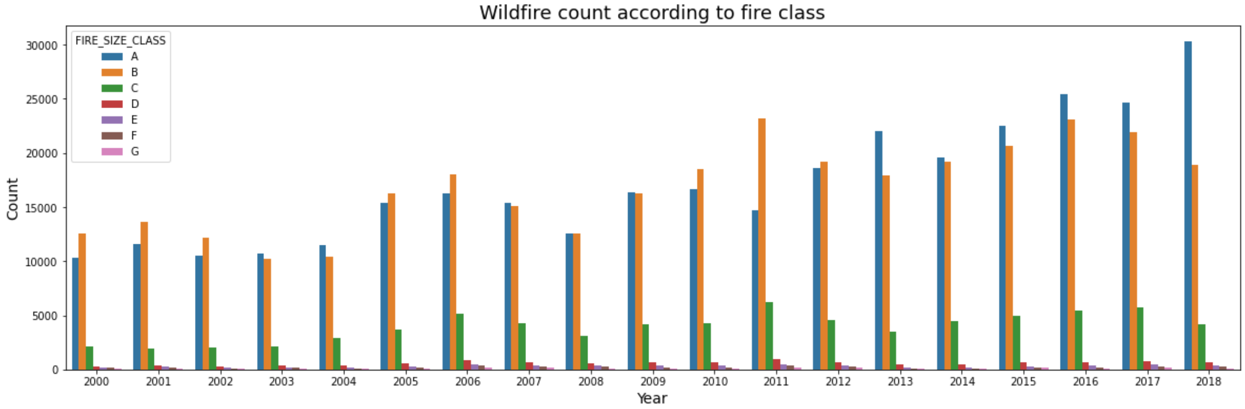

*Figure 2: Wildfire Count per Year by Class*

### Massachusetts

As we concentrated on Massachusetts fire cases, we noticed that most wildfires incidents occurred in recent years, particularly in 2016 and 2017. Among all the wildfire size classes, the majority of wildfire cases were classes A and B (Figure 3). We also observed that Berkshire County, Hampden County, Worcester County, Barnstable County, Essex County, Dukes County, Suffolk County, Middlesex County, Hampshire

County, and Norfolk County had wildfires from 2000 to 2018. Among those counties, Essex County has the most wildfire incidents (Section 4.1.2).

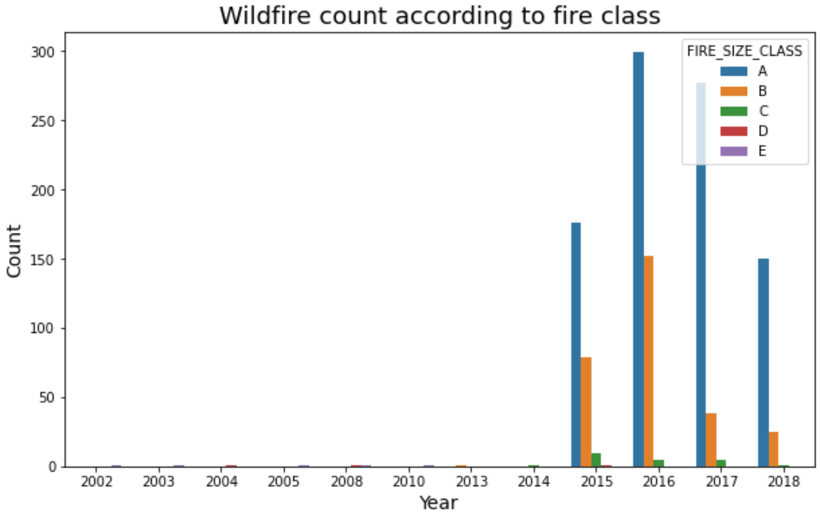

*Figure 3: Wildfire Count according to Class in MA*

## Maine

After conducting some preliminary analysis we found that the years with the most wildfires in the state of Maine were 2001 and 2016 with 963 and 658 wildfires respectively. The most common fire size class was Class A, which was closely followed by Class B (Figure 4). The counties in Maine that experienced the most wildfires since 2000 were Penobscot County and Aroostook County (Section 4.1.3).

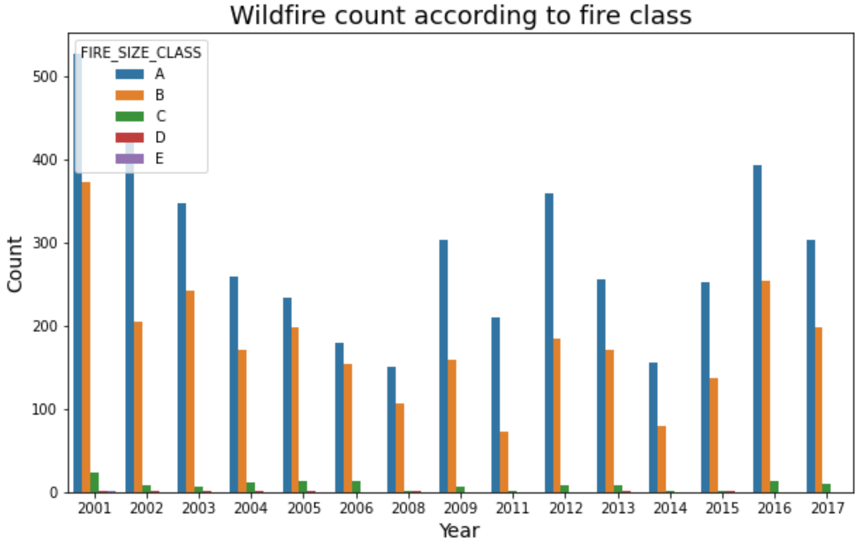

*Figure 4: Wildfire Count according to Class in MA*

## Connecticut

Connecticut displays similar patterns as Massachusetts. To begin with, we noticed that most wildfires incidents occurred in recent years, particularly in 2016 and 2017. In Connecticut, the most common fire class was Class A, similar to both MA and ME (Figure 5). Furthermore, the top county that experienced the most wildfires in CT was Litchfield County (Section 4.1.4).

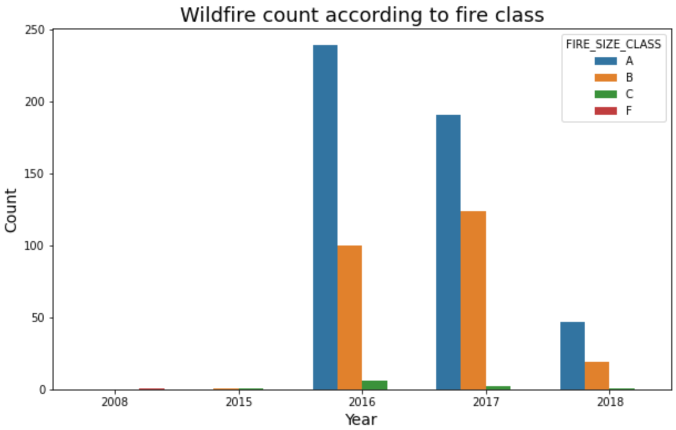

*Figure 5: Wildfire Count according to Class in MA*

## Drought dataset

For the drought dataset, we took drought data for each county in the New England region from 01/01/2000 to 01/01/2020. The drought dataset contains the Drought Severity and Coverage Index(DSCI). The DSCI is an index for converting drought levels from the U.S. Drought Monitor map to a single value for an area. DSCI values are part of the U.S. Drought Monitor data tables. Possible values of the DSCI are from 0 to 500. Zero means that none of the areas is abnormally dry or in drought, and 500 means that all of the areas are in D4, exceptional drought.

We first calculated the average DSCI in Massachusetts,  Maine, and Connecticut for the past 10 years. We observed that the average DSCI in these three states displays a similar pattern. These three states all relatively have high DSCI values in 2001, 2002, 2016, and 2017 (Figure 6).

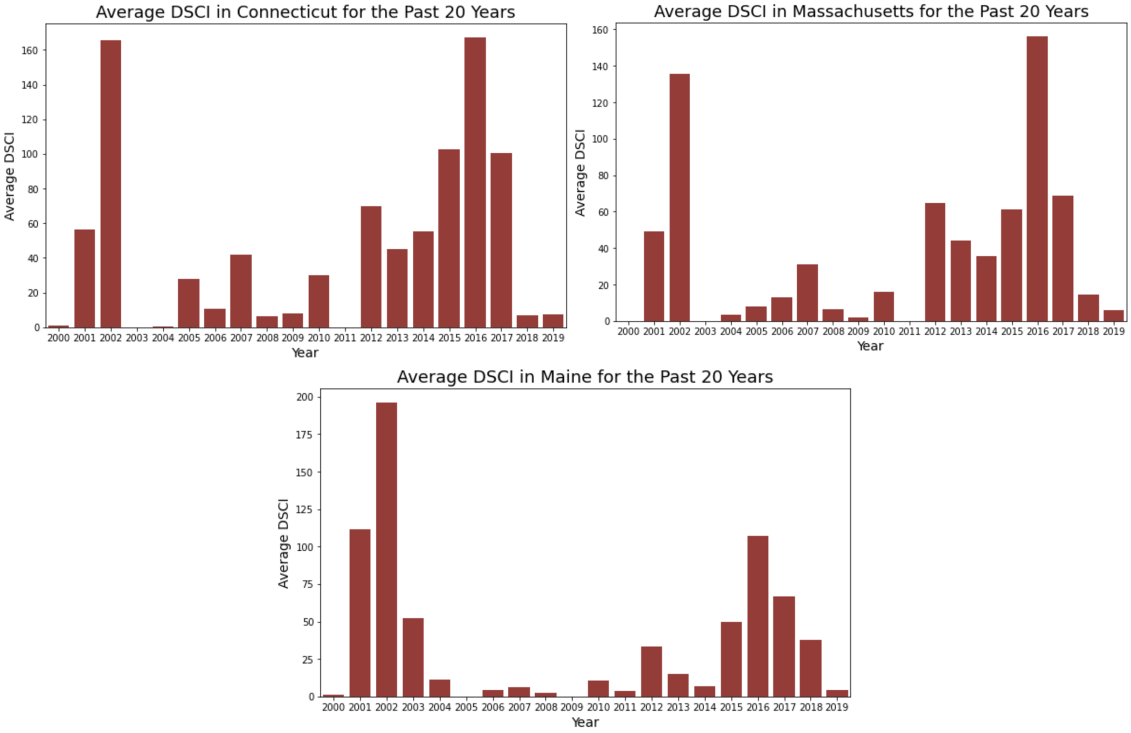

*Figure 6: Average DSCI per Year for CT, MA, and ME*

We then calculated the average DSCI according to the county. York County and Plymouth County have the highest average DSCI in the New England region. We then plotted the DSCI of York County and Plymouth County according to the year. We noticed that similar to the average DSCI of three states, these two counties share the same trends in terms of DSCI. The average DSCI value in Massachusetts, Maine, and Connecticut are 39.2, 35.05, and 44.14 respectively. (Section 4.2.1).

### Census dataset

We first found county median household income and household income per capita according to counties. This was followed by finding the counties with the lowest and highest of the above-mentioned values in order to observe if certain demographic attributes differed between these counties.

We found that Nantucket County and Fairfield County have the highest median income and income per capita respectively. Furthermore, we can see that Somerset County and Piscataquis County have the lowest income per capita and median income, respectively. Next, we decided to illustrate a correlation matrix to identify certain variables we wanted to further explore. We decided to further explore poverty rates, household numbers, population, and county according to race percentages (Section 7.1.1).

We observed that there is a relationship between the aforementioned variables and the percentage of races in each county. Firstly, we noticed that in general the higher percentage of whites present in a county the smaller the population and therefore also the lower the number of households. This relationship is inverted for the other races which illustrates that as the percentage of the other races increases in the counties so does the population and the number of households. We can also observe that counties in New England are predominantly lived in by white people. When looking at other graphs, we can see that Suffolk County has higher percentages of Black or African Americans, Hispanic or Latinos, and Asians. We can also see that American Indian and Alaskan Natives predominantly reside in Washington County, whereas Native Hawaiian and Other Pacific Islanders, as well as Hispanic or Latinos, predominantly reside in Providence County (Section 7.1.1).

## Feature Engineering

Our feature engineering encompassed merging the wildfire and drought datasets together in a meaningful way for machine learning use. The base table we used was the wildfire dataset as the goal of our machine learning model was to predict the severity of wildfires in the New England region. Therefore, we needed to merge the two datasets in a way that for every date for which a wildfire occurred in a specific county there was a respective DSCI value associated with the row. We firstly confirmed that both datasets contained the same counties as these along with the dates would be the metrics we'd be joining the tables on. Next, we created a function that was able to calculate the average DSCI score of the previous *n* weeks (Section 5.3). Finally, we merged the dataset by matching the state, county, and date of each dataset and appending 3 separate columns with averages of the previous 4, 8, and 12 weeks. The final dataset included the original wildfire dataset columns with an additional 3 DSCI average columns.

## Machine Learning

### Preparing the Dataset for Machine Learning

To start with machine learning with the data we had collected, the first step was to subset the data available to us on a particular day before the event occurred. We included features such as Latitude and Longitude of the counties, FIPS code, drought indexes, etc. To prepare the data before using any algorithm, we used dummy variables for our categorical variables, such as County, State, FIPS Code, etc., using one-hot encoding. Finally, we were able to split the data into train and test sets for training and evaluating purposes with a split of ⅓ into test data and the rest as training data.

## Implementing Models

We approached this problem with two different target variables, fire size, and fire class. The problem was to use regression models on the data set and predict the values for fire size, i.e., given certain conditions, how big will the fire be? We started with simple linear regression, trying to minimize the MSE by getting a straight line on the data set. We evaluated the model using two metrics, MSE and MAE, to decide which model to use. MAE was useful in our case as we had a lot of outliers in our dataset, so MSE will not give us an accurate understanding of model performance. The other model we decided to use was Random Forest Regressor, a tree-based ensemble method. It makes a bootstrap sample and uses random selection on decision nodes, making it helpful in determining the important features of our dataset. We used multiple parameters to find the best fit for our model (Section 6.1).

Later, we approached our second problem classification, i.e., trying to predict a specific fire class given certain conditions. We used multiple classification models to predict the fire class accurately. The fire class was divided into six categories - A to F, depending on its impact. Before making any classification, we first checked the distribution of data. It was highly imbalanced as more than 95% of rows belong to classes A and B. We used logistic regression with class = 'balanced' parameter to approach this problem, which provides calculated weights to every class while calculating the probability of belonging to a class. This helped us improve our ROC-AUC score. The next model we tried was Random Forest Classifier, an ensemble technique to predict the classes. We trained this model after validating multiple hyperparameters to optimize the ROC-AUC score. The last model we used was a Balanced Random Forest Classifier from the imblearn library. It makes random undersampling of the dominant class to make the model better learn about the data. Both the models worked for us getting the predictions for both fire size and fire class (Section 6.2).

## Results

| Regression Model | MSE | MAE |
|---|---|---|
| Linear Regression | 168.83 | 3.091 |
| Random Forest | 272.94 | 2.179 |

*Table 1: Regression Model Results*

| Classification Model | Accuracy | F1 Score |
|---|---|---|
| Logistic | 0.582 | 0.588 |
| Random Forest | 0.627 | 0.611 |

| | | |
|---|---|---|
| Balanced RFC | 0.254 | 0.265 |

*Table 2: Classification Model Results*

## Cost Calculation

After some exploratory data analysis looking at the census data, we conducted research in regards to the costs of wildfires. We found that on top of the cost to the fire department, households also incur both direct and indirect costs. Households pay for homeowners insurance that covers wildfire costs among others. Wildfire costs tend to include: smoke damage, burnt materials, soot damage, chemical damage, water damage, and structural damage. The focus of our project is the different impacts of a wildfire between richer versus underserved counties. To associate a number and be able to analyze this relationship we decided to calculate the cost per household of a wildfire.

After research, we decided to take into account the deductible that is part of the insurance plan and the increase in insurance premium after a fire claim has been submitted as a function of the value of the housing unit. This cost is an estimate of what may need to be paid out of pocket by a household that experiences a wildfire. The deductible is paid by default if the cost is larger and we decided to also add the increase in premium as the increase in yearly premium is a direct effect of the wildfire. In this case, we are only taking into account the following year's increase in insurance premiums that needs to be paid. We are using the household value as a basis for calculating the insurance premium as one should select household insurance where the dwelling is approximately the household value.

The outcome of these calculations is a specific estimated cost of a single wildfire occurring in one year. The median household income was then divided by the calculated cost value to indicate the proportion of income the cost makes up. After doing these calculations for each county we looked at the counties that illustrated the highest and lowest proportions and analyzed various attributes like poverty level, the race percentages present, and the number of households in each county (Section 7.2). The following formulas were used to calculate the aforementioned values:

*Wildfire Cost Estimate: (AVG_RATE*RATE INCREASE % AFTER FIRE) + DEDUCTIBLE*
*Proportion of Wildfire Cost = MEDIAN HOUSEHOLD INCOME/ESTIMATED WILDFIRE COST*

# Key Findings

In regards to our machine learning models, we used different algorithms to predict fire sizes and fire classes. In terms of the regression model we achieved the highest accuracy with a random forest regressor. This model resulted in an MAE of 2.179 which indicates that on average, the model predicted the fire size with approximately 2 acres more or less than the target size. Considering that this MAE is lower than that found with linear regression we can see that the random forest model was successful. When looking at the classification model we see a similar outcome. The random forest classifier performed the best with an accuracy of approximately 62%. This accuracy score, although seemingly low, actually represents a successful model as the baseline accuracy is approximately 17%. Considering that the severity of wildfires can be attributed to fuel, weather, and topography, and our model only takes into account partial weather information, namely drought, there is extensive opportunity to improve model prediction by gaining more data in regard to the attributing factors of wildfire severity. This would aid in more accurate predictions and therefore more valuable models that can be used in mitigation strategies.

From the census data, we know the counties with more low-income groups should be protected more by the government. First, African American, Asians and other ethnic minorities have fewer households on average, and their income is relatively low. Usually, low-income groups live in low-lying, less protected areas in counties and also have less resources to handle natural disasters (The Nature Conservancy, n.d.). These are the places that are most vulnerable to wildfire in the first place. Second, after calculating the proportions we found that in the county where wildfire costs make up the highest proportion (Lamoille County, Vermont) of median household income have a lower percentage of Hispanic or Latinos, Black or African American, Asian and a larger population of white residents. The opposite was the case for the county where wildfire cost represented the lowest proportion of median household income (New London County, Connecticut). These findings should be taken into consideration when implementing wildfire mitigation strategies as we can clearly see the discrepancy between the effects of wildfires on households.

We also decided to look more narrowly at counties (Section 7.3). We found that New Haven County has the highest average DSCI over the past 20 years, however it does not appear in the counties in New England that experience the most wildfires. In fact the top 10 counties with the highest average DSCI over the past 20 years do not seem to have experienced the highest number of wildfires. This may indicate that drought does not accurately represent a predicting factor in wildfires. We can also see that the counties where wildfire cost represents the largest proportion of median household income are not the counties that have experienced the largest number of wildfires in New England. Another insight we gained from the data is that 5 of the top 10 counties with the highest poverty rates are also present in the counties with the top 10 number of wildfires in the past 20 years. This illustrates that many wildfires occur in poor counties and thus may require a larger aid from the government and organizations.

| County | Poverty Rate, percent | Number of Wildfires (Position in Relation to all New England) |
|---|---|---|
| Aroostook County, ME | 15.3 | 840 (2nd) |
| Piscataquis County, ME | 14.6 | 409 (10th) |
| Somerset County, ME | 14.5 | 509 (6th) |
| Oxford County, ME | 13.0 | 503 (7th) |
| Penobscot County, ME | 12.1 | 1011 (1st) |

*Table 3: Poverty Rate vs Wildfire Occurrence by County*

Some of these findings are surprising as they are not what we initially expected. To begin with, we expected that counties where the proportion of wildfire cost represented more income compared to other counties would have mostly non-white residents. The opposite was found. Furthermore, we also saw that counties with the highest average DSCI are not the same counties that experience the most wildfires. This was unexpected as we hypothesized drought to be a relevant predictor for wildfires. This relationship was also visible in the machine learning models as they were performing poorly due to the fact that the predictor data did not include enough wildfire severity explanatory variables like fuel and topography. One finding that did corroborate our hypotheses was that counties that experienced the most wildfires were also the one with the highest poverty rates, which was confirmed by the census data.

# Business Impact and Implications

The business implications of our project can be categorized into the following categories:
1. Insurance - Insurance Sector deals with buildings affected by a wildfire. We have calculated the cost impact of a wildfire in a certain county, which is directly proportional to the claims to be paid by insurance companies. They will have an idea to calculate premiums with respect to the county a certain building falls into as some counties are more prone to wildfires given the geographical conditions. Also, insurance firms can research more into what is the proportion of claims they are paying for damages by wildfires and compare it with the revenue from premiums received by households to gain a better understanding to calculate the cost they are sustaining.

2. FEMA - The US Fire Department is working on a fire-safe America. Our project is helping them identify the parameters which are highly correlated with the fire size. These factors can be further researched to minimize the fire instances. Suppose, a certain drought average leads to a wildfire, they can declare that zone as camping free, as a lot of campfires leads to a wildfire. Also, we analyzed demographics in counties. There is a vast difference in counties in terms of mean income per capita, which makes poor counties such as Somerset County, Maine, Piscataquis County, Maine etc. more vulnerable to wildfires when compared with rich counties such as Nantucket County, Massachusetts, Fairfield County, Connecticut, etc. FEMA can focus on these counties to ensure that they are able to meet the cost of wildfire.

3. Damage Limitation - The model is able to predict wildfires on a certain date of the year, which gives an opportunity to limit the damage by taking necessary steps in those areas. If the predicted size is large then the area can be evacuated and the firefighters can be prepared for an unfortunate event like this and try to stop it before the fire increases. The counties with low income per capita and high frequency of wildfires can prevent additional damages because of delayed discovery of wildfire. This will help these counties to save more money.

# Limitations

There are two major limitations in this project that should be addressed in future research. The first one is the limited access to data. After we finalized our project topic, we started to look for datasets regarding wildfires, drought, and demographic. We faced the problem of having limited access to these datasets really quickly. In order to gain access to these datasets, we have to first know who manages and has access to the dataset. We then tried to reach out to the organization or specific person who had access to the data. The process is time-consuming and sometimes we do not get any reply or permission at all. For our future analysis, if we could gain access to the potentially useful research datasets for our research, our model and our results might be more precise and reliable.

The second limitation concerns the insufficient data. After cleaning the wildfire dataset, it contained 741,000 wildfire cases. However, after keeping only the New England region wildfires as this is the region this project focused on, the dataset only contained 9,630 wildfire cases. Insufficient wildfire data might affect our machine learning model result. For future analysis, if we could get more wildfire cases data in the New England region, our machine learning classification result will be more precise. Furthermore, if we could get a dataset that not only presents wildfire occurrences but also shows the conditions that don't cause a wildfire, we can probably predict the likelihood of a wildfire event in an area.

We believe that more wildfire, drought, and census data will definitely make more accurate projections to help counties better prepare to prevent fire damage. As our research primarily focuses on the New England region, we hope that this research can be the benchmark in the future, applying to other parts of the United States that are not only experiencing increasing drought and wildfires conditions but also have large disparities between the communities where resources are distributed unevenly. We also hope that U.S. fire departments and counties can develop more detailed and complete plans and better preventive measures based on our research to deal with the possible occurrence of droughts and fires, reduce the damage and losses caused by fires and provide residents with more reasonable compensation and assistance.

## Bibliography

Davies, I. P., Haugo, R. D., Robertson, J. C., Levin, P. S. (2018). The unequal vulnerability of communities of color to wildfire. PLOS ONE, 13(11). https://doi.org/10.1371/journal.pone.0205825

Moore, A. (2021, December 3). Explainer: How wildfires start and spread. College of Natural Resources News. Retrieved April 30, 2022, from https://cnr.ncsu.edu/news/2021/12/explainer-how-wildfires-start-and-spread/

Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5

The Nature Conservancy. (n.d.). New research shows increased wildfire risk among minorities. The Nature Conservancy. Retrieved May 8, 2022, from https://www.nature.org/en-us/about-us/where-we-work/united-states/washington/stories-in-washington/wildfires-impact-minorities/

U.S. Drought Monitor. (n.d.). Drought severity and coverage index. Drought Severity and Coverage Index | U.S. Drought Monitor. Retrieved February 25, 2022, from https://droughtmonitor.unl.edu/DmData/DataDownload/DSCI.aspx

## Tools and Resources

| **Resource** | **Link** |
|---|---|
| GitHub Repository | https://github.com/kratik28/capstone_b8.git |
| Code Notebook | https://colab.research.google.com/drive/1kNl4dbwig7_pV2ryk3jV9U6R-tipjuvY?usp=sharing |
| Google Drive Dataset Folder | https://drive.google.com/drive/folders/13Sf7qz21NGKORMfex8GIobWiHAAMKJAx?usp=sharing |
| Final Presentation | https://docs.google.com/presentation/d/1xSgohxnEbrCFJgD8PSljv9qUsDzDd9emS1FHllQkKtE/edit?usp=sharing |
| Final Poster | https://www.canva.com/design/DAE-5AVDh6Q/cFQRoIzAqkG8WErb7aA4CA/view?utm_content=DAE-5AVDh6Q&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton |