U.S. Wildfire Analysis

Problem Statement

Wildfires are a significant environmental issue that also endanger human lives. It is a complex issue with varying variables in play as wildfires occur both naturally and unnaturally. Our goal is to recommend preventive solutions by leveraging the National Interagency Fire Center's (NIFC) data that spans 1985 to 2020.

Background on NIFC

NIFC is home to the national fire management programs of each federal fire agency, along with partners including the National Association of State Foresters, the U.S. Fire Administration, and the National Weather Service. Fire management under this larger umbrella is designed to achieve not only suppression goals, but to accomplish a broad spectrum of natural resource objectives, and do so in an efficient, cost-effective manner.

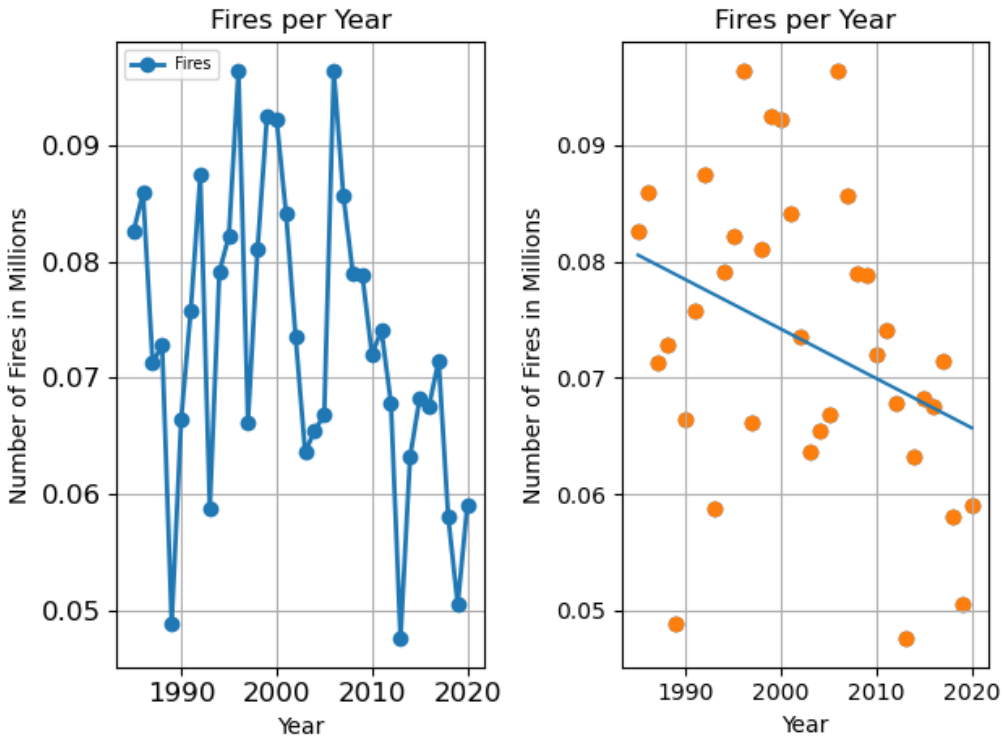
Data Wrangling

The raw dataset from Kaggle via NIFC is a simpler one and one that is complete as is with no nan or missing values. As a result, data cleanup was minimal, but it focused on:

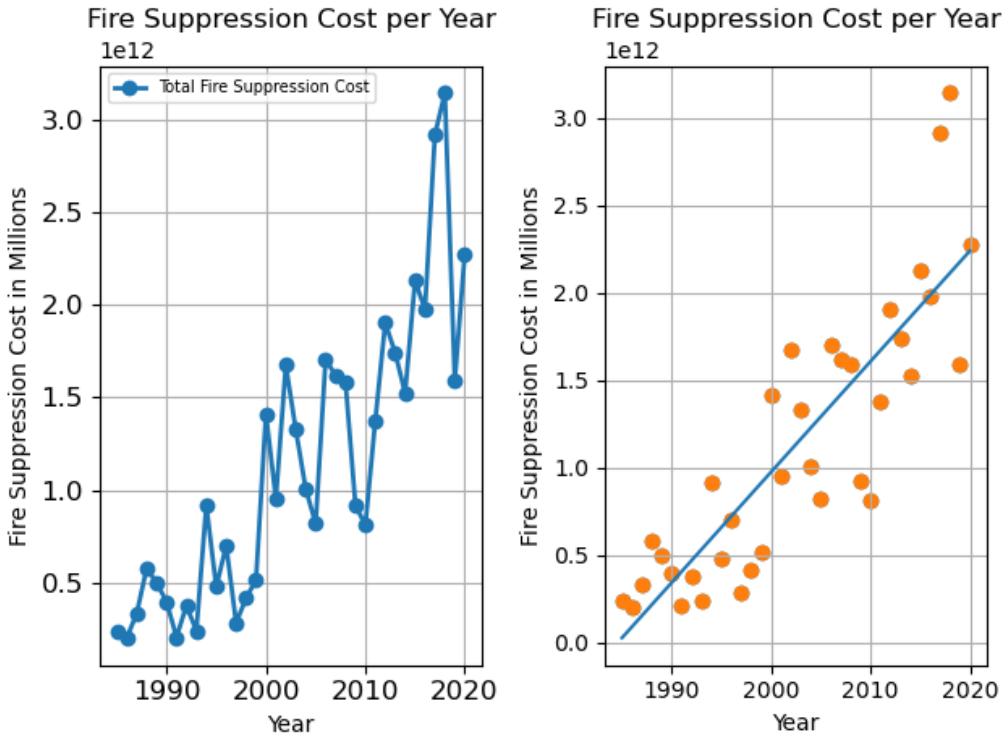
- Stripped commas and USD format of applicable fields.
- Added currency and country fields for better interpretability.
- Renamed fields for better interpretability.
- Converted data types to integer format since applicable numeric values are object format due to their USD format.
- Scaled all numeric values to same unit for ease of analysis (in Millions).
- Built in additional fields to capture key metrics.
    - Acres Burnt per Fire
    - Suppression Cost per Fire
    - Year over Year Change in Fire Suppression Cost

Exploratory Data Analysis (EDA)

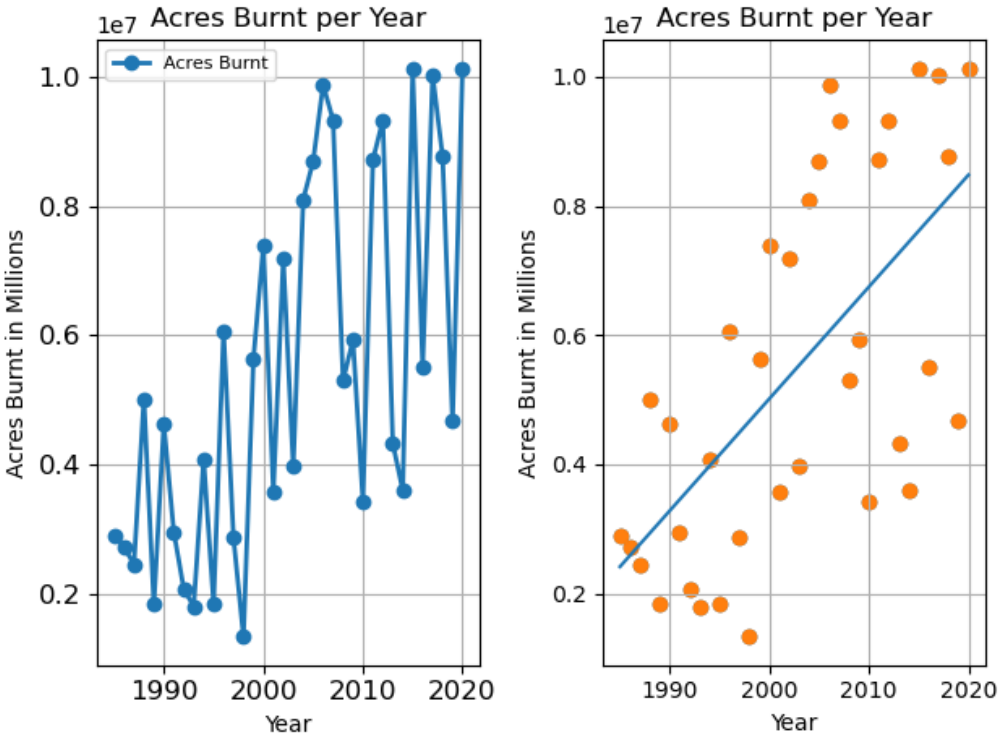
Digging into the data, almost 148 million acres have been burnt over the past two decades compared to around 48 million acres the prior 15 years. Further, suppression costs have exceeded $34 billion over the past two decades compared to over $6 Billion over the prior 15 years. The time periods may not reflect the same duration year wise, but the trends are very concerning, nonetheless.



Outliers are distinguished at years 1989, 1996, 2006, 2012, and 2019 and there is a good bit of variability in the grouping.



Outliers are distinguished at years 2017 and 2018 but the grouping is tight.



Outliers are not overly distinguished as there is a good bit of variability in the grouping.

We see that over time, the number of fires has declined while both the acres burnt, and the total fire suppression cost have increased. This is a key finding for my clients and one that requires further investigation beyond this report. A number of explanations could factor in. Also, there are a few outlier data points (called out under graphs above) but they were not discarded.

The dataset has nine features which is not many, but the correlations are positively strong enough to leverage. Two features, “Country” and “Currency” are not numeric. Hence, their exclusion from the heat map above. To no surprise, there is a relatively strong positive relationship between fields like "Suppression Cost per Fire" and "Acres Burnt" (correlation of 0.64) and there is a strong positive relationship between fields like "Total Fire Suppression Cost" and "Suppression Cost per Fire" (correlation of 0.96). Inversely, there are low positive correlations and low negative correlations for the feature "Fires." The feature "Year over Year Change in Fire Suppression Cost" also has low positive correlations. These findings are valuable for modeling and predicting.

Modeling

The dataset lends itself to supervised machine learning and specifically, regression analysis. With that in mind, I focused on Random Forest Regression, KNN Regression, Decision

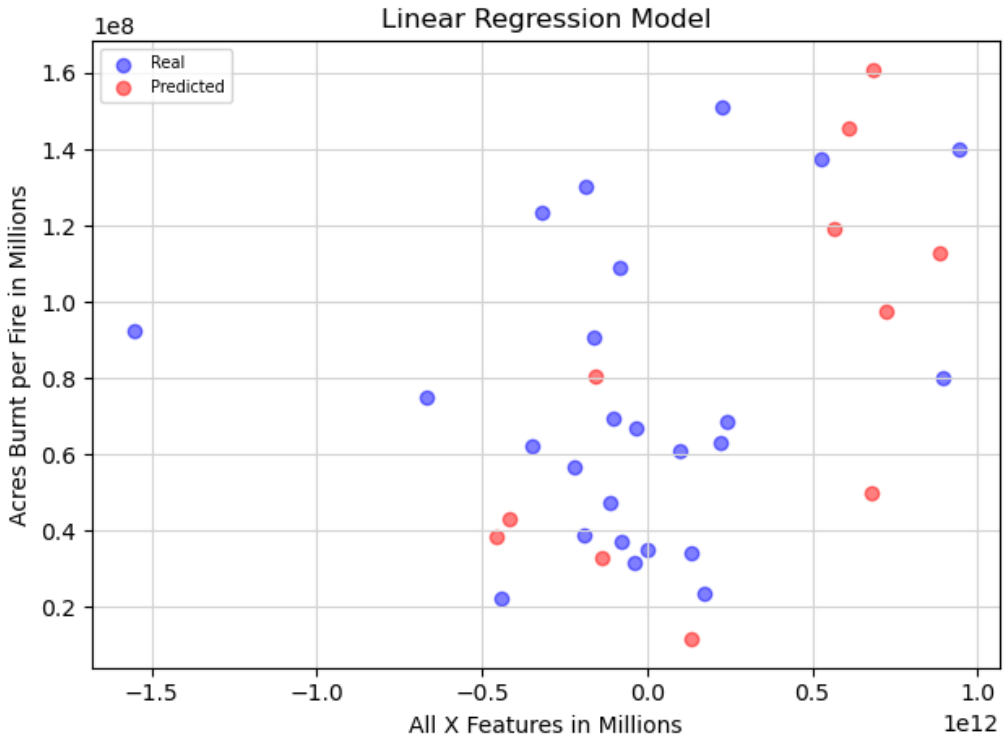
Tree Regression, and Linear Regression. I dropped features “Years,” “Country,” and “Currency.” The “Years” feature was dropped as regression analysis’ basis is not predicting time, so I did not see value in its inclusion. The target variable was the feature “Acres Burnt per Fire,” and I allocated a 30%/70% test, train split on the dataset as well. Below are the model results:

Below are model performance results:

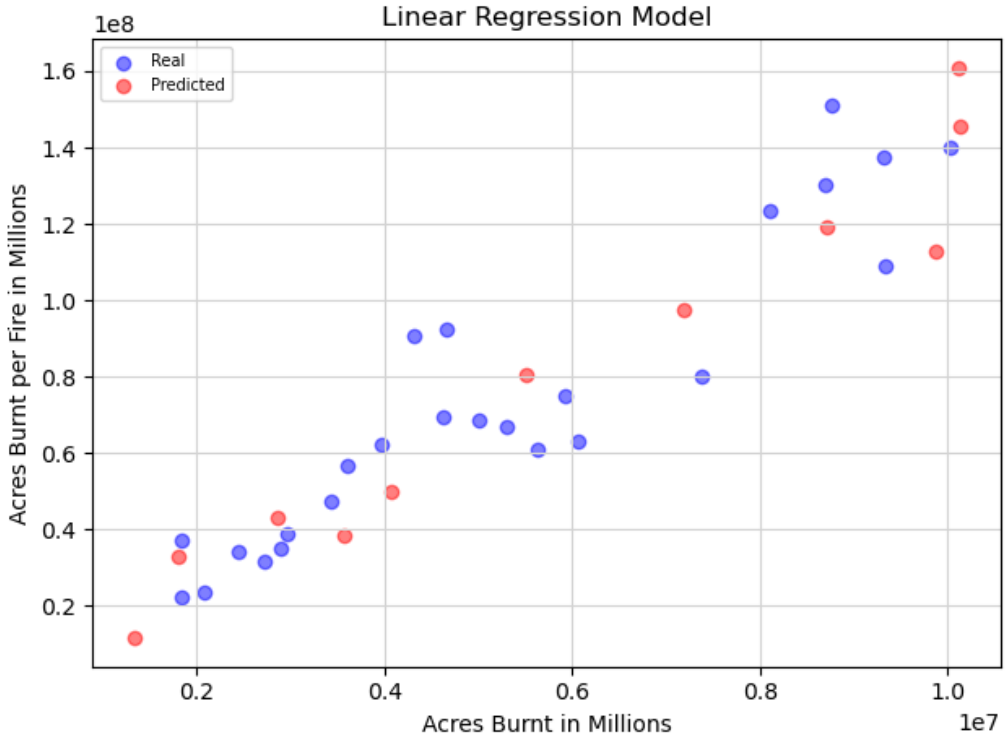
1. Random Forest (model_1)
   - Root Mean Squared Error (RMSE): 16,103,675.23
   - Mean Absolute Percentage Error (MAPE): 0.2147
   - R Squared: 88.73%
2. KNN Regression (model_2)
   - Root Mean Squared Error (RMSE): 30,032,137.55
   - Mean Absolute Percentage Error (MAPE): 0.3979
   - R Squared: 60.81%
3. Decision Tree (model_3)
   - Root Mean Squared Error (RMSE): 20,615,589.49
   - Mean Absolute Percentage Error (MAPE): 0.2598
   - R Squared: 81.53%
4. Linear Regression (model_4)
   - Root Mean Squared Error (RMSE): 5,053,806.5
   - Mean Absolute Percentage Error (MAPE): 0.0635
   - R Squared: 98.89%

I focused on RMSE, MAPE, and R Squared as my metrics to give me a more thorough understanding of the quality of the respective model performances. RMSE determines the absolute fit of the model to the data as it measures the average difference between a model's predicted values and its actual values. MAPE considers the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors. For both RMSE and MAPE, the lower the value the better. Finally, R Squared measures the proportion of variance in the dependent variable that can be explained by the independent variable. The greater the value the better. Vaguely, 0.65/65% or greater is considered pretty good.

The Linear Regression model clearly performed the best, so I leveraged it for my predictive analysis seen below.



The model performs well at predicting the "Acres Burnt per Fire" by using all features tied to my X variable.



By isolating the X variable to one feature, "Acres Burnt," the model performs very well at predicting Acres Burnt per Fire over time.

Recommendations

- Push local governments for more conversation resources and practices to be implemented to protect forests.
- Make the public more aware of their potential impact on this crisis. Not all wildfires are naturally occurring.
- Increase staff capacities across conservation agencies and first responders to better handle wildfires so they can be put out timelier.

Future Work

As mentioned in the EDA section, there could be a number of explanations behind why acres burnt, and total fire suppression cost have increased over time while the number of fires has decreased. Below are future considerations:

- Leverage an additional dataset or two focused on global CO2 levels to see if they factor into trends found in the Wildfire dataset.
    - Dive into avenues that can explain why wildfires are most potent and costly even though the number of them over the years have declined.
- Explore impact of inflation over time to better flatten out analysis of cost features.
- Explore which regions or states are most impacted by wildfires.