Katelyn Wong
Dr. Siddharth Vishwanath
Math 189
3/16/2025

Analyzing Forest Fire Severity in California

1. Introduction

Wildfires pose a significant threat to ecosystems, economies, and human lives, especially in California, where climate conditions and human activity make wildfires a recurring disaster. In January 2025, major wildfires swept through Los Angeles, destroying tens of thousands of acres and causing economic losses estimated between $250 billion and $275 billion[1]. Given the severe impact of such events, understanding the factors that influence fire severity is critical for prevention and mitigation strategies. This study aims to analyze historical wildfire and weather data to investigate which combination of meteorological factors (temperature, pressure, humidity, wind speed, wind direction) most strongly predict wildfire severity in the United States, as measured by acres burned between 2012 and 2015. We hope this will provide insights to help predict future fire risk.

2. Data Sources and Processing:

I utilized two primary datasets for this project:

A. Kaggle: Historical Hourly Weather Data 2012-2017

While I discovered this dataset on Kaggle, the original data was acquired using the Weather API on the OpenWeatherMap website. The weather dataset contains seven individual datasets that each represent an important meteorological attribute: city attributes, humidity, pressure, temperature, weather description, wind direction and wind speed. The city_attributes.csv file contains latitude and longitude information on 36 different cities across the U.S., Canada and Israel.

	City	Country	Latitude	Longitude
0	Vancouver	Canada	49.249660	-123.119339
1	Portland	United States	45.523449	-122.676208
2	San Francisco	United States	37.774929	-122.419418
3	Seattle	United States	47.606209	-122.332069
4	Los Angeles	United States	34.052231	-118.243683

Fig. 1: Snapshot of city_attributes.csv

The remaining 6 datasets each contain five years' of hourly measurements, each with 45252 observations and 37 variables. Of the 37 variables, one is the datetime measure while the other 36 correspond to the weather conditions in each city at a specific time.

	datetime	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	Las Vegas	Phoenix	Albuquerque	...
1	2012-10-01 13:00:00	0.0	0.0	150.0	0.0	0.0	0.0	0.0	10.0	360.0	...
2	2012-10-01 14:00:00	6.0	4.0	147.0	2.0	0.0	0.0	8.0	9.0	360.0	...
3	2012-10-01 15:00:00	20.0	18.0	141.0	10.0	0.0	0.0	23.0	9.0	360.0	...
4	2012-10-01 16:00:00	34.0	31.0	135.0	17.0	0.0	0.0	37.0	9.0	360.0	...
5	2012-10-01 17:00:00	47.0	44.0	129.0	24.0	0.0	0.0	51.0	8.0	360.0	...

Fig. 2: Snapshot of wind_direction.csv

B. Kaggle: 1.88 Million US Wildfires

The wildfire dataset is a spatial SQLite database containing 27 individual datasets. It was originally generated to support the now-retired Fire Program Analysis (FPA) program. The database includes 1.88 million records of wildfires that occurred in the United States from 1992 to 2015. Given the volume of data, I focused on the Fires dataset, which contains information collected from U.S. federal, state, and local reporting systems.

My first dataset includes weather information across the U.S., Canada, and Israel, while the second is limited to U.S. states. Additionally, the first dataset covers the period from 2012 to 2017, whereas the second spans from 1992 to 2015. Considering the size of the Fires dataset, I began by randomly sampling 1,800 rows. I then merged the two datasets based on their overlapping time frame and geographic location.

2.1) Data processing, merging and subsetting:

I have two separately generated datasets, so in order to find relationships between multiple variables (columns) across them, I first needed to merge the datasets based on certain variables. Specifically, the weather dataset contains weather conditions in specific cities over time, while the fire dataset contains fire records at specific latitude and longitude coordinates over time. Based on similarities in location and time, I was able to merge the two datasets.

I began by extracting the cities in the weather dataset and retrieving their corresponding latitude and longitude coordinates. Then, for each fire (row) in the fire dataset subset and for each city in the weather dataset, I calculated the great-circle distance using the haversine distance method. I set a threshold of 100 miles to determine whether a fire and city were close enough to be considered

related. If the distance was less than 100 miles and the timestamps matched, I added a row to the merged dataframe containing both the fire record and the corresponding city's weather conditions.

Since the full fire dataset contains 1.8 million records, I initially used a subset to accelerate the merging process during pipeline development. I randomly sampled 1,800 rows from the fire dataset, and the merging process took approximately 31.5 seconds to run on a MacBook.

2.2) Correlation matrix (heatmap):

Among the many variables in the fire dataset, I selected FIRE_SIZE_CLASS for prediction purposes. The variable FIRE_SIZE_CLASS_encoded is derived from the standard FIRE_SIZE_CLASS definition, where A represents the least severe fires and G the most severe. Note that in the final dataset, only fire sizes A, B, and C are present due to subsetting and merging.

The weather condition variables I used in the correlation analysis are: Humidity, Wind_Direction, Temperature, Pressure, and Wind_Speed.

3. Exploratory Data Analysis (EDA)

3.1) Missingness Analysis

I began by exploring missing values in my dataset.

	Missing Values	Percentage (%)
ICS_209_INCIDENT_NUMBER	1669	100.000000
ICS_209_NAME	1669	100.000000
MTBS_ID	1669	100.000000
MTBS_FIRE_NAME	1669	100.000000
COMPLEX_NAME	1669	100.000000
FIRE_CODE	1524	91.312163
LOCAL_FIRE_REPORT_ID	1475	88.376273
CONT_TIME	499	29.898143
DISCOVERY_TIME	372	22.288796
FIRE_NAME	350	20.970641
CONT_DATE	350	20.970641
CONT_DOY	350	20.970641
COUNTY	219	13.121630
FIPS_CODE	219	13.121630
FIPS_NAME	219	13.121630
LOCAL_INCIDENT_ID	72	4.313960

Fig. 3

As shown in Figure 1, 16 out of the 53 columns in my merged dataset contain missing values. Upon further investigation, I found that all 16 of these columns are of the object data type. Importantly, there were no missing values in the key weather-related numeric columns, which I planned to use for analysis.

I conducted a missingness analysis on the columns with missing values. Several of these columns—such as FIRE_CODE (91.3% missing) and LOCAL_FIRE_REPORT_ID (88.3% missing)—appear to be missing at random (MAR). Data is considered MAR when the likelihood of

missingness depends on other observed variables rather than the values of the data itself. In this case, the missingness could be related to the state or reporting unit, as some states or units may be less consistent in reporting certain fields.

On the other hand, columns like FIRE_NAME are likely not missing at random (NMAR). For example, smaller fires that were extinguished quickly may have been recorded but never given a name. Data is considered NMAR when the missingness is related to the unobserved value itself, rather than other variables.

3.2) Outlier detection

Next, I utilized Z-scores to check my dataset for outliers. I used a threshold of 3, such that values with z-scores greater than 3 or less than -3 were considered outliers. I did not have any negative values in my dataset, and only investigated the numeric columns for outliers.

```
Outliers in each column (Z-score > 3):
 Humidity                  0
Wind_Direction             0
Temperature                0
Pressure                  22
Wind_Speed                30
FIRE_SIZE_CLASS_encoded    0
dtype: int64
```

Fig. 4

As shown in Figure 2, most of the columns contained values that were reasonable. Pressure and Wind_Speed appeared to have 22 and 30 values respectively that were outliers. I visualized these using a boxplot to understand them better.

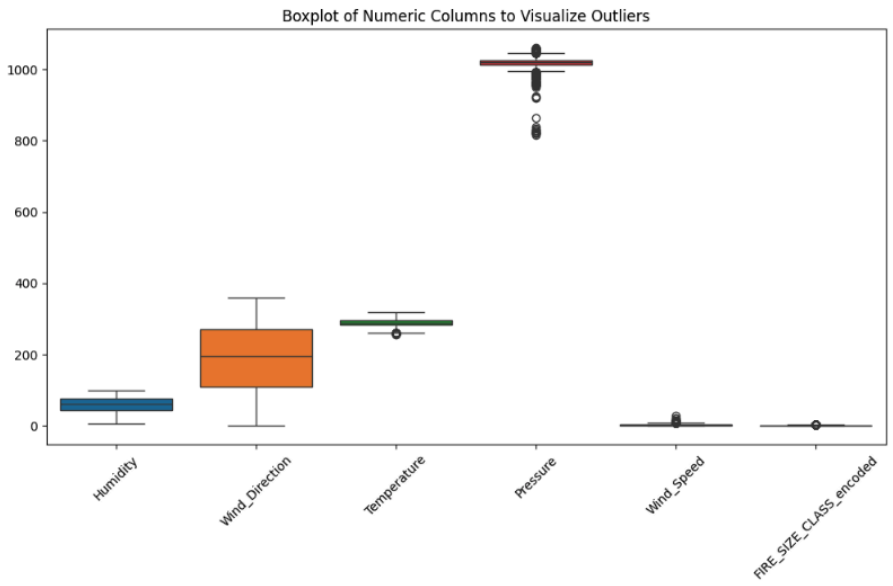

Fig. 5

Wind_Direction appears to have a wide range of values, evidenced by the long whiskers. Pressure, Wind_Speed and Temperature seem to contain outliers, given the individual circles on and outside of the whiskers. However, according to Temperature's z-score, all its values are in a reasonable range. It is understandable for weather attributes to have extreme values. They're an important part of the story that helps highlight the cause behind fires, so I am not eliminating the outliers.

3.3) Choropleth Map showing varying fire sizes across the country

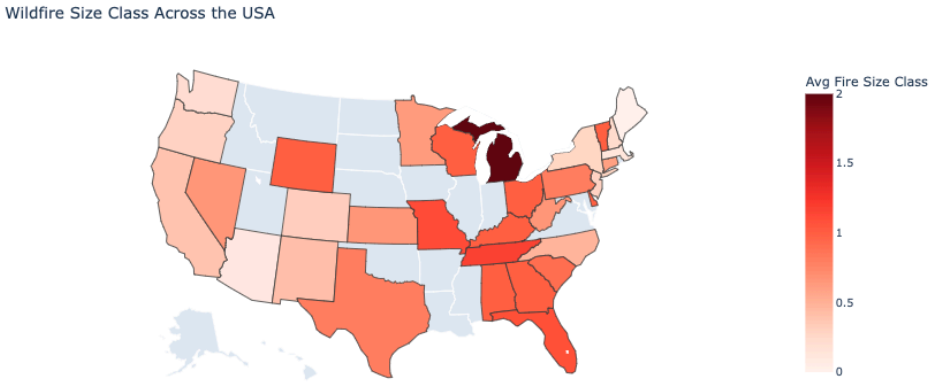

Fig. 6

After investigating the outliers, I was curious to see how the sizes of fires varied across the United States. The 'FIRE_SIZE_CLASS_encoded' column has 3 categories that fires are classified into depending on the size of the fires, with the 3 most common being:

0: (0, 0.25] acres i.e, fires that are greater than 0 but less than or equal to 0.25 acres

1: [0.26-9.9] acres

2: [10.0-99.9] acres

I built a choropleth map that shows the average size of the fires per state across the country using these 3 fire categories. There are 21 U.S. states represented in my dataset, and Figure 4 portrays the average fire sizes in those states. I chose a continuous color scale to emphasize how Florida, for instance, has larger fires on average than Colorado.

3.4) Bubble chart

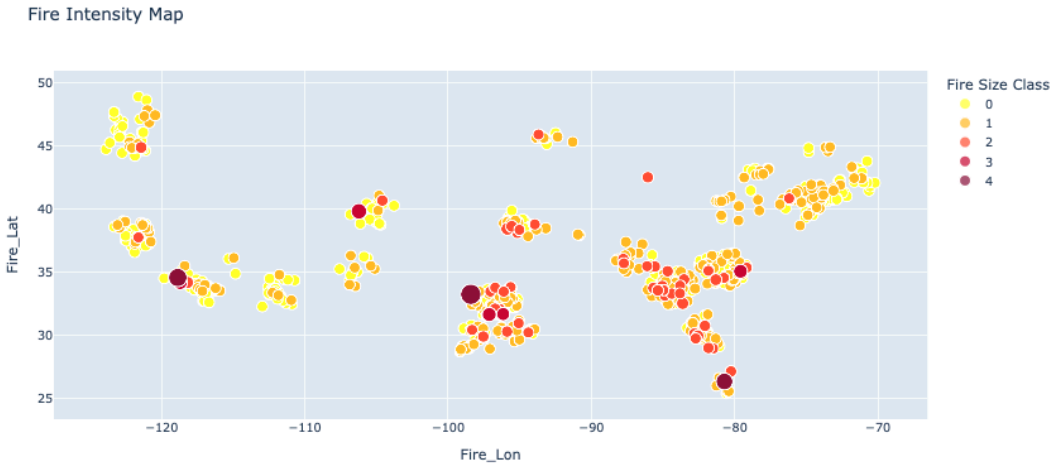

Fig. 7

Similar to the U.S. map in Figure 4, I wanted to create a map that would portray how the fire sizes vary by latitude and longitude. I ensured the color choices for the legend would align with reader expectations, and so I chose yellow, orange and red to portray the increasing fire sizes.

4. Statistical Analyses and Modeling

Each weather parameter dataset was first reshaped using a melt operation to standardize the format. These reshaped datasets were then merged on ‘Datetime’ and ‘City’. I computed a correlation matrix, which helped us identify significant relationships among the weather variables.

4.1) Logistic Model Building

In my initial prediction model, I randomly sampled from my dataset (n = 20,000) and employed an ordinal logistic regression model (Proportional Odds Model) to predict fire size, an ordinal variable with three levels, based on various weather and location features (all continuous variables). Specifically, I modeled fire size (FIRE_SIZE_CLASS_encoded) as a function of humidity, wind direction, wind speed, temperature, pressure, wind, latitude (City_Lat), and longitude (City_Lon).

Before constructing the regression model, I observed that certain predictor variables exhibited high Variance Inflation Factors (VIFs), indicating multicollinearity issues. To address this, I standardized these variables, which successfully mitigated the multicollinearity problem. However, I chose not to standardize latitude and longitude, despite their higher VIFs, because their interpretability relies on maintaining their original scale.

Instead of standardizing longitude, I opted to remove it from my feature set. This decision was based on the reasoning that latitude is generally more predictive of temperature, as cities closer to the

equator tend to have higher temperatures on average. By retaining latitude and excluding longitude, I aimed to reduce redundancy while preserving meaningful geographic information.

After standardizing the remaining variables and removing longitude, the updated predictor variables have the following corresponding VIFs:

predictor variables:

	Variable	VIF
0	Humidity_standardized	1.087349
1	Temperature_standardized	1.151861
2	Pressure_standardized	1.116165
3	Wind_Direction	4.562096
4	Wind_Speed	2.920621
5	City_Lat	5.033826

Fig. 9

At this point, I ran the ordered logistic regression and obtained the following results:

```
Optimization terminated successfully.
         Current function value: 0.893174
         Iterations: 30
         Function evaluations: 34
         Gradient evaluations: 34
                             OrderedModel Results
==============================================================================
Dep. Variable:     FIRE_SIZE_CLASS_encoded   Log-Likelihood:                -17410.
Model:                        OrderedModel   AIC:                         3.484e+04
Method:                Maximum Likelihood   BIC:                         3.492e+04
Date:                   Sat, 15 Mar 2025
Time:                           23:33:34
No. Observations:                   19492
Df Residuals:                       19482
Df Model:                               6
==============================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Humidity_standardized         -0.0592      0.015     -3.960      0.000      -0.089      -0.030
Temperature_standardized      -0.2472      0.016    -15.559      0.000      -0.278      -0.216
Pressure_standardized          0.1881      0.017     10.944      0.000       0.154       0.222
Wind_Direction                -0.0007      0.000     -4.709      0.000      -0.001      -0.000
Wind_Speed                     0.0590      0.007      7.907      0.000       0.044       0.074
City_Lat                      -0.1216      0.003    -35.244      0.000      -0.128      -0.115
0.0/1.0                       -4.3367      0.126    -34.489      0.000      -4.583      -4.090
1.0/2.0                        0.9553      0.011     88.595      0.000       0.934       0.976
2.0/3.0                        0.7283      0.032     22.655      0.000       0.665       0.791
3.0/4.0                       -0.0127      0.093     -0.136      0.892      -0.195       0.169
==============================================================================
```

Fig. 10

4.2) Interpretation of Model Results

A. Lower humidity is associated with larger fire sizes:

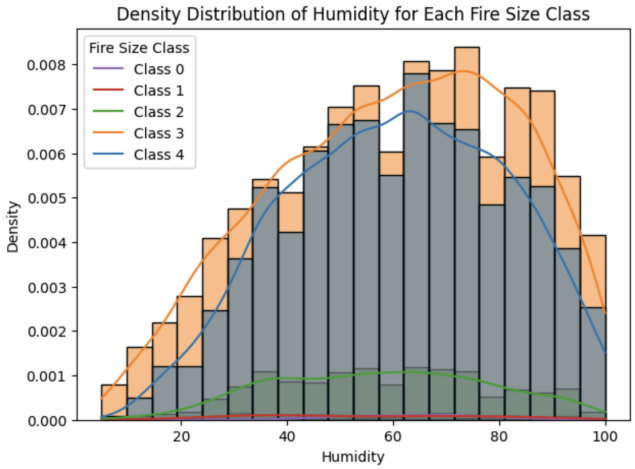

Fig. 11

Holding all other variables constant, a one standard deviation increase in humidity likely decreases the log odds of being in a larger fire size category by 0.0592 ($p < 0.001$ indicates a strong effect).

a. These results align with scientific observations as lower humidity levels indicate drier air, which can lead to drier environmental conditions that can lead to rapid fire spread.

B. Lower temperature is associated with larger fire sizes:

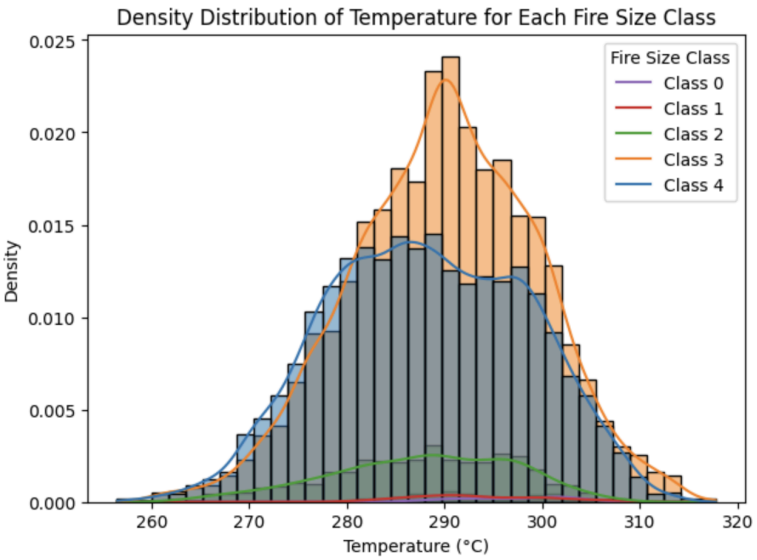

Fig. 12

Holding all other variables constant, a one standard deviation increase in temperature likely decreases the log odds of being in a larger fire size category by 0.2472 ($p < 0.001$ indicates a strong effect).

a. While this may seem counterintuitive, temperature is likely correlated with other environmental variables that directly affect fire spread, resulting in omitted variable bias. For example, cold temperatures are correlated with colder seasons and dry conditions, which could promote larger fires. In addition, in certain areas, higher temperatures may be associated with humid, fire-suppressing conditions, which could reduce fire spread and size. This could be a potential area for further research.

C. Higher pressure is associated with larger fire sizes: Holding all other variables constant, a one standard deviation increase in pressure likely increases the log odds of being in a larger fire size category by 0.1881 ($p < 0.001$ indicates a strong effect).

a. According to scientific observations, higher atmospheric pressure is generally associated with low humidity, clear skies, and prolonged dry spells/heat waves, leading to conditions that also promote the start and spread of wildfires.

D. Higher wind speed is associated with larger fire sizes:

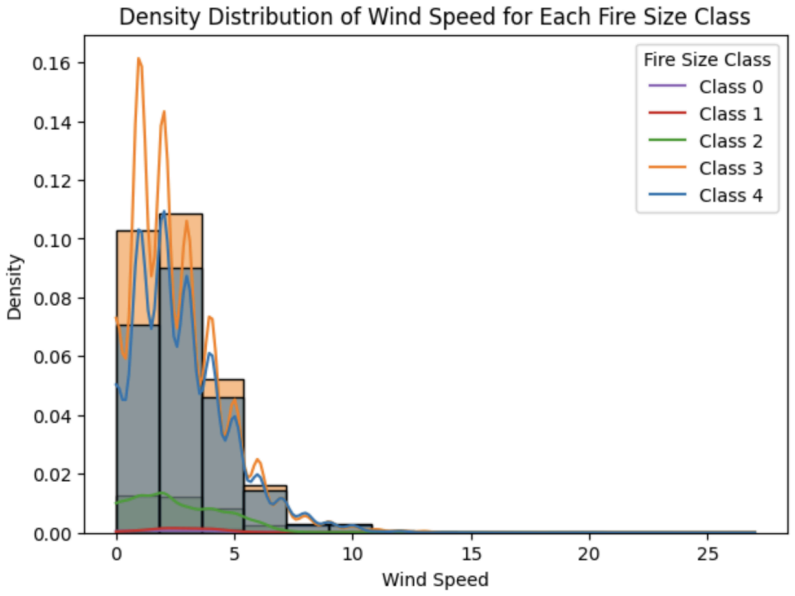

Fig. 13

Holding all other variables constant, a one unit increase in wind speed likely increases the log odds of being in a larger fire size category by 0.0590 ($p < 0.001$ indicates a strong effect).

a. These results also align with scientific observations. Because wind supplies fresh oxygen to fire, wildfires in higher wind conditions tend to have higher intensity and combustion efficiency. In addition, stronger winds can direct flames horizontally, leading to faster spread of fires.

E. Wind direction has a significant but very small effect on larger fire sizes: Holding all other variables constant, a one unit increase in wind direction likely decreases the log odds of being in a larger fire size category by 0.0007 (likely not a significant prediction feature for fire size).

F. Locations with lower latitudes are associated with larger fire sizes: Holding all other variables constant, a one degree decrease in latitude likely increases the log odds of being in a larger fire size category by 0.1216 ($p < 0.001$ indicates a strong effect).
    a. Because latitude is strongly correlated with temperature and regions with lower latitudes are closer to the equator and receive more direct sunlight year round, they may have more frequent droughts and lower humidity levels, leading to conditions that are prone to large wildfires.

G. Threshold (Cut-Off) Values Interpretation
    a. 0/1: The log-odds of a fire being in size class 1 or higher compared to class 0 is -4.3367 ($p < 0.001$). Because this value is highly negative, it suggests that class 0 fires are very common, and the probability of moving to class 1 or higher is relatively low.
    b. 1/2: The log-odds of being in size class 2 or higher compared to class 1 is 0.5553 ($p < 0.001$). Since this value is greater than the 0/1 threshold, it indicates that transitioning from class 1 to class 2 is easier than moving from class 0 to class 1.
    c. 2/3: The log-odds of being in size class 3 or higher compared to class 2 is 0.7283 ($p < 0.001$). This suggests that the likelihood of moving into class 3 continues to increase as fire size grows.
    d. 3/4: The log-odds of being in size class 4 or higher compared to class 3 is -0.0127 ($p = 0.892$), indicating an insignificant effect. Since this value is close to zero and the p-value is above the 0.05 significance level, it suggests that the model struggles to distinguish between class 3 and class 4 fires. Even under extreme weather conditions, the transition between these two largest fire sizes may be less clearly defined in the dataset.

4.3) Ordinal Logistic Regression Model Summary:

The model identifies several significant predictors of fire size, with lower humidity, lower temperature, higher pressure, higher wind speed, and lower city latitude all increasing the odds of a larger fire. While wind direction is statistically significant, its small coefficient suggests a negligible practical effect on fire size. Additionally, an analysis of threshold values indicates that smaller fire sizes are more common, but as fire size increases, the model is more likely to classify fires into higher size categories, reflecting a natural progression in fire growth.

5. Model Experimentation with Machine Learning (Random Forest Classifier and Neural Networks)

After experimenting with the logistic regression model, I tried to use other machine learning models for predicting the fire.

5.1) Random Forest Classification

First I selected the Random Forest model, which is a traditional and interpretable machine learning model based on decision trees. I have a standard 8:2 split of training and testing set, and run on 18000 samples that have 6 fire size categories.

Fig. 14

I achieved a solid prediction accuracy of approximately 71% using the random forest model. I then examined the model's feature importance and found that longitude was the most important variable, followed by humidity. This result aligns well with expectations. As shown in Figure 7, wildfires are clustered geographically, with the most significant fires occurring in regions like Florida and Texas, which have distinct longitudes compared to other parts of the country. Additionally, humidity is well known to be closely related to fire behavior, making its high importance consistent with prior understanding.

```
print("Accuracy:", model.score(X_test, y_test))
Accuracy: 0.7058220056424724
```

Fig. 15

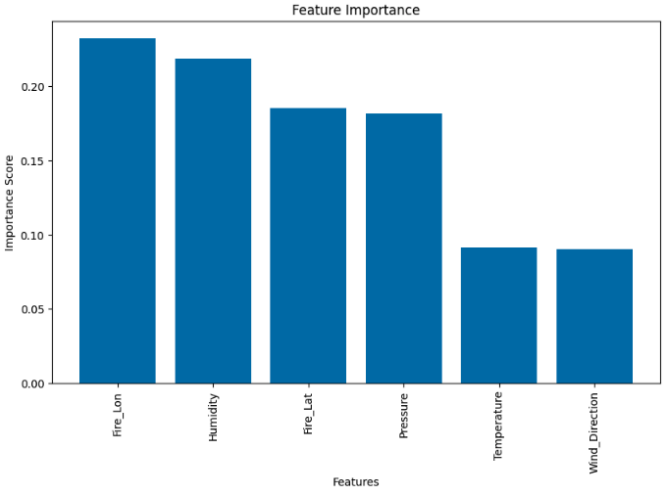

Fig. 16

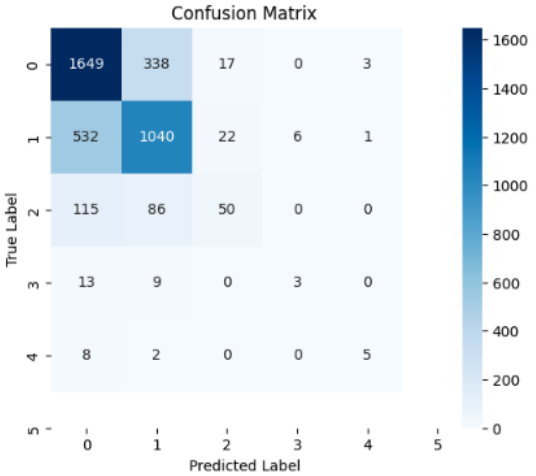

Fig. 17

5.2) Neural Networks

Secondly, I tried to train a more complicated neural network model using the same data as the random forest model. It is a 5-layer fully connected network, adding Batch Normalization and Dropout for each layer separately.

```
Epoch 10: Train Loss = 1.2441, Val Loss = 1.1968
Epoch 20: Train Loss = 1.1136, Val Loss = 1.0815
Epoch 30: Train Loss = 1.0934, Val Loss = 1.0632
Epoch 40: Train Loss = 1.0364, Val Loss = 1.0221
Epoch 50: Train Loss = 1.0457, Val Loss = 1.0494
Epoch 60: Train Loss = 1.0589, Val Loss = 1.0005
Epoch 70: Train Loss = 1.0213, Val Loss = 0.9808
Epoch 80: Train Loss = 0.9939, Val Loss = 0.9620
Epoch 90: Train Loss = 1.0088, Val Loss = 0.9842
Early stopping triggered
<All keys matched successfully>
```

Fig. 18

Training loss:

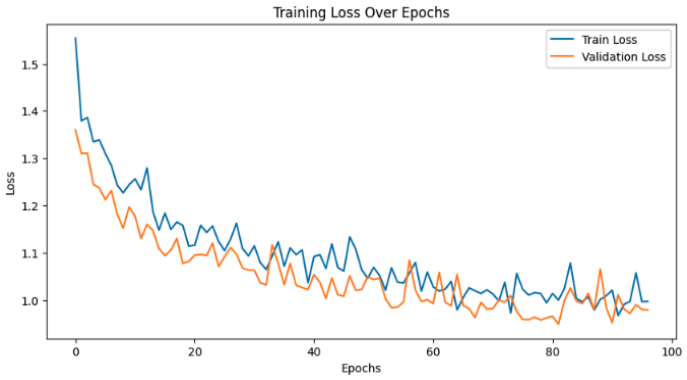

Fig. 19

However, this more complicated model did not perform well compared to the random forest model. From the confusion matrix, I can see that the model did not catch the skewed data distribution (majority of the data has fire level 0), and predicted a lot of false positives.

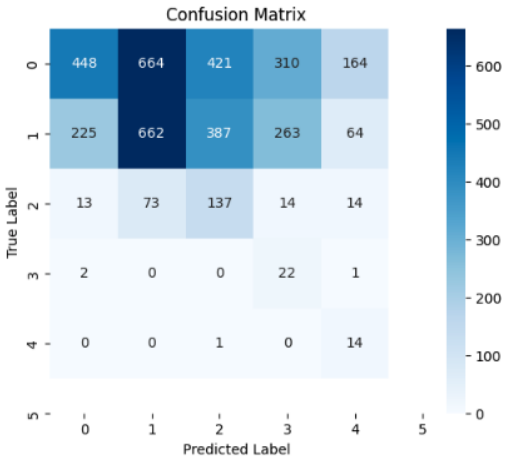

Fig. 20

6. Limitations and Future Work

A. Incomplete Data: My data set is limited to fire information in 21 U.S. states. This means my predictions may not generalize well to other states.

B. Temporal Predictions: My model has been trained on data from 2012 - 2015. However, the wildfire patterns observed in the historical data may not hold true for future predictions due to climate change. This means the model has to be constantly updated with real-time weather and wildfire data to ensure it results in reliable predictions.

C. Omitted Variable Bias: Because my data set is limited to certain predictor variables, my model may suffer from omitted variable bias, meaning that certain key factors that influence

wildfire size are not included in the analysis due to the lack of sufficient wildfire data. One major example of this is temperature interactions; I included temperature as a standalone predictor, but temperature generally interacts with other environmental conditions, such as fuel load, lightning strikes, and land cover type. As such, the omission of these variables could introduce bias in my predictions, making my estimates less reliable in certain conditions.

D. Class Imbalance: My dataset is the disproportionate representation of small fires (Class 0) compared to larger fires (Classes 3 and 4). As a result, the model may become biased toward predicting smaller fires accurately while underperforming in identifying and classifying the relatively rare large fires. This imbalance can lead to skewed evaluation metrics and reduce the model's real-world applicability.
E. Outliers: Pressure and Wind speed have more outliers than Humidity, Wind direction, or Temperature.

7. Conclusion

In this study, I investigated the relationship between meteorological variables and wildfire severity using a dataset that merges wildfire records (2012–2015) with weather observations from the United States. Humidity and temperature emerged as key environmental factors, with low humidity and, unexpectedly, lower temperatures increasing the odds of larger wildfires. Wind speed and pressure were also associated with larger fires, suggesting that windy, high-pressure weather systems may accelerate fire spread. Additionally, regions closer to the equator tended to experience larger fires, likely due to generally warmer and drier climates.

This project has several limitations. Expanding the dataset to include additional states, more recent years, and other environmental variables—such as fuel moisture, topography, and land use—could improve predictive accuracy. I also anticipate that continued advances in machine learning models, along with the integration of domain-specific data (e.g., lightning strikes), will enhance model performance and practical utility.

The results of this project provide a preliminary framework for identifying high-risk conditions that contribute to large wildfires. These insights can help inform more effective resource allocation and targeted prevention efforts in areas and under weather conditions most prone to severe fire events.

Citations

Danielle, & Monica. (2025, January 16). AccuWeather estimates more than $250 billion in damages and economic loss from LA wildfires. AccuWeather. https://www.accuweather.com/en/weather-news/accuweather-estimates-more-than-250-billion-in-damages-and-economic-loss-from-la-wildfires/1733821

Selfish Gene. (2018). Historical hourly weather data [Dataset]. Kaggle. https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

Tatman, R. (2020). 188 million US wildfires [Dataset]. Kaggle. https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires