Analysis of U.S. Wildfires

Chanheum Park
CSCI 5502
Boulder, Colorado
Chanheum.Park@colorado.edu

Feelgyun Oh
CSCI 5502
Boulder, Colorado
Feelgyun.Oh@colorado.edu

Taeho Kim
CSCI 5502
Boulder, Colorado
Taeho.Kim@colorado.edu

KEYWORDS

datasets, data mining, wildfire, machine learning, deep learning

ACM Reference Format:
Chanheum Park, Feelgyun Oh, and Taeho Kim. 2020. Analysis of U.S. Wildfires. In *Boulder '20: Data Mining, Oct 07–09, 2020, Boulder, CO*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recently a pair of wildfires in northern Colorado set records as the largest and second largest wildfire in Colorado state history burning almost 630 square miles of land destroying over 200 residential structures. Not only does such fires cause a huge financial detriment to the community and environment, but they also cause mental issues to does who had direct impact from these disasters. Wildfires have a negative economic impact on communities by making recreation and tourism unappealing, and affecting agricultural production. Local communities often become concerned about the effects of smoke on health and safety as well.

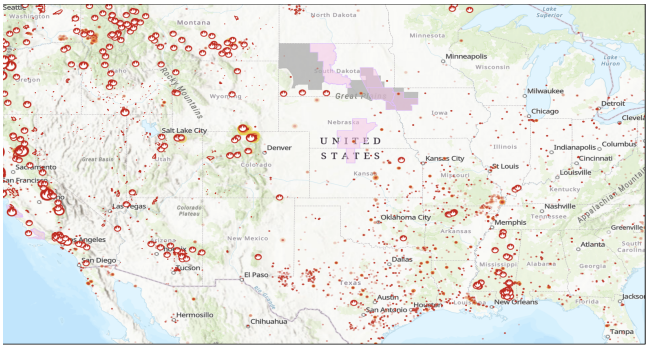


Figure 1: US Wildfire

The increasing concern over wildfires has induced a re-emphasis by the federal government to stop this trend. To effectively prevent wildfires, it seems crucial to have a sound understanding of the patterns and causes of wildfires [2].

Boulder '20, Oct 07–09, 2020, Boulder, CO

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this project, we analyze the U.S. wildfires data from Kaggle [8], and find various climate datasets from different government institutions, and extract features from them. With the selected features, we visualize various causes of wildfires, how big they were, frequently occurring regions, and so on. We then use several Machine Learning skills to make predictions on wildfires occurrence and damage to our economy.

The goals of this project are to visualize wildfire maps, predict some characteristics, and find patterns from a dataset produced by the national Fire Program Analysis (FPA) system. The dataset consists of many attributes with some we think has the potential of providing us answers to our questions as follows:

- Has the occurrence of wildfire increased over time?
- Which factor contributes most to causing wildfires?
- What is the relationship between wildfire and weather?
- Are there any noticeable wildfire patterns depending on the region (state)?
- Can we provide a model that can predict the economic burden caused by wildfires?

In this paper, we first describe the recent approaches and trials to analyze wildfires. Next, we introduce the dataset and data preprocessing in Section 3. In Section 4, we analyze US wildfires using a variety of data mining techniques, such as classification, clustering, and regression, as well as considering data visualization. In Section 5, we evaluate our prediction and data mining analysis. Lastly, we summarize the overall contents of this paper and provide the prospect of future research.

2 RELATED WORK

Many researches have been conducted throughout the past decades to analyze the causes of wildfires and predict them to improve fire prevention. One of the early studies shows that the dry fuelbeds are needed to enable successful ignition and spread. In addition, quantities, structure and moisture contents of the fuels that consist of the fuelbeds are one of the key factors to the spread of wildfires and its damage [6].

Other research that focus on the human-caused wildfire identifies an interesting result. It shows that cigarettes consumed and adult smoking rates are associated with the wildfires occurrences. In addition, the number of reported arson wildfires on national forests are closely followed changes to downward trend in the rate of all crimes; violent index crimes, and nonviolent property crimes [6].

There is also strong evidence that regional warming and drying play a role in the increasing fire frequency. However, human caused wildfires represented 84% of the 1.5 million wildfires from 1992 to 2012 included in the analysis of the research and has tripled the length of the wildfire season [4]. With California having the highest occurrence of wildfire, research showed that human-caused fires

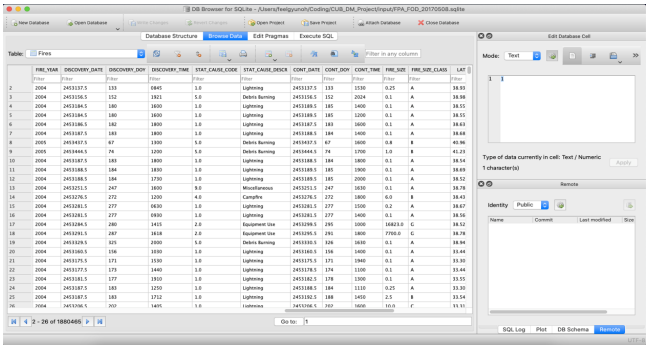


Figure 2: SQLite database

dominated the fire records and showed a positive correlation with population density for the first two thirds of the record, but then showed a decline in recent decades.

Research in Australia shows that some ignition sources with a small limited number like electrical distribution lines results a much larger size of wildfire [5]. Another research conducted in California supports this by showing that electrical distribution lines was one of the top two causes of area burned [7].

Prestemon et al. [6] suggests a conceptual model that shows linked ignitions in societal, biophysical, fire prevention and management variations that illustrates the complexity of understanding the relationship between sources and how they change over time.

Not only do these researches try to provide an in-depth analysis of wildfire data, but they also provide an insight on how to prevent and minimize economic loss due to wildfires, but even with such efforts, more needs to be done to expand our understanding on wildfire prevention.

3 DATA PREPROCESSING

3.1 Wildfires Dataset

The dataset is referred to as the Fire Program Analysis fire occurrence database (FPA FOD) which includes 1.88 million geo-referenced wildfire records from 1992 to 2015. These records were acquired from the reporting systems of federal, state, and local fire organizations. This dataset is a SQLite database with a vast variety of attributes, as shown in Figure 2. The original dataset has multiple features and sparse characteristics. We need to remove any biased data and interpolate values for some empty cells. Feature selection removes meaningless data, so it will help us focus on our task with important factors. Due to the substantial amount data and the large pool of attributes, it is important to use different attributes accordingly to the needs of analysis in the data mining process. Among them, we select several features as follows:

- **FOD_ID**: Global unique identifier to distinguish each wildfire case
- **FIRE_YEAR**: Calendar year in which the fire was discovered to check the historical trend of wildfires
- **MONTH** & **DAY_OF_WEEK**: Extracted features from date on which the fire was discovered (Julian Date ⟶ Gregorian format)
- **FIRE_SIZE**: Estimated acres that are damaged by wildfire

```
In [9]: df = pd.read_sql_query("SELECT LATITUDE, LONGITUDE, STATE, FIPS_CODE, FIPS_NAME FROM 'Fires'", cnx)
        df.head(10)
```

Out[9]:

	LATITUDE	LONGITUDE	STATE	FIPS_CODE	FIPS_NAME
0	40.036944	-121.005833	CA	063	Plumas
1	38.933056	-120.404444	CA	061	Placer
2	38.984167	-120.735556	CA	017	El Dorado
3	38.559167	-119.913333	CA	003	Alpine
4	38.559167	-119.933056	CA	003	Alpine
5	38.635278	-120.103611	CA	005	Amador
6	38.688333	-120.153333	CA	017	El Dorado
7	40.968056	-122.433889	CA	None	None
8	41.233611	-122.283333	CA	None	None
9	38.548333	-120.149167	CA	005	Amador

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="http")
location = geolocator.reverse("40.968056, -122.433889")
print(location.address)
```

```
Cascade Wonderland Highway, Vollmers, Shasta County, California
```

Figure 3: Processing missing values

- **FIRE_SIZE_CLASS**: Classifying damaged acres [A (0-0.25 acre), B (0.26-9.9), C (10.0-99.9), D (100-299), E(300-399), F(1000-4999), G (5000+)]
- **LATITUDE** & **LONGITUDE**: Visualization and checking the location
- **STATE**: Two-letter alphabetic code for the state
- **FIPS_CODE**: Three-digit code from the Federal Information Process Standards (FPS)
- **DURATION**: Wildfire duration (CONT_TIME - DISCOVERY_TIME), [CONT_TIME: Time of day that the fire was declared contained, DISCOVERY_TIME: Date on which the fire was discovered]
- **LABEL**: Code for the (statistical) cause of the fire
- **STAT_CAUSE_DESCR**: Description of the (statistical) cause of the fire

Originally, the dataset contains Julian date which is the number of days elapsed from the start of the cycle to a specific date in 7,980 years. We convert Julian dates into Gregorian format to create new

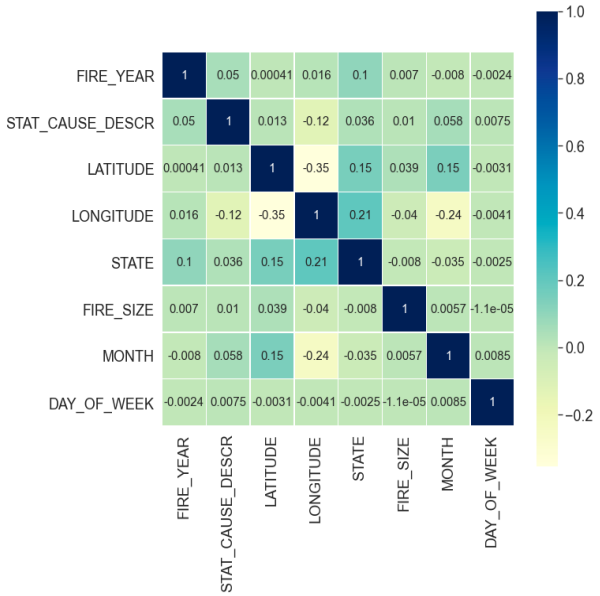
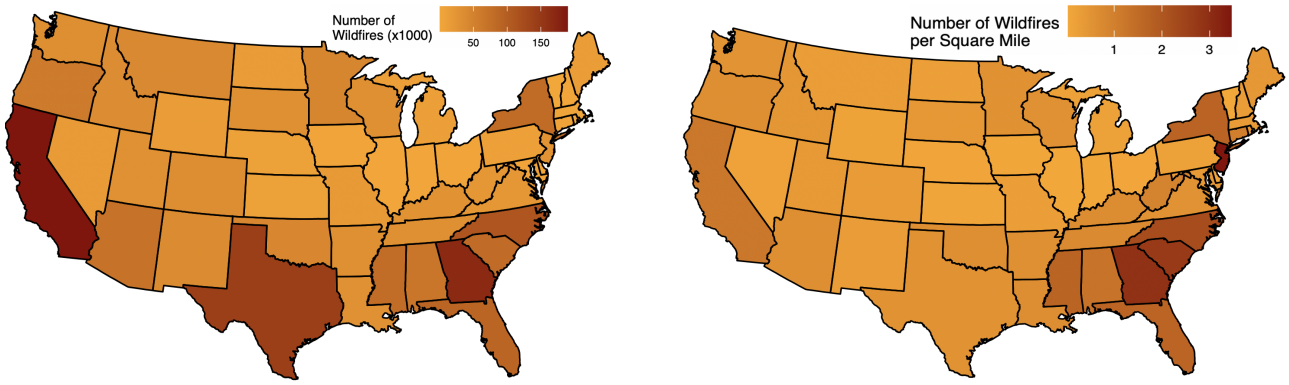


Figure 4: Selected Features' Correlation Matrix



(a) Not normalized **(b) Normalized by area**

Figure 5: Number of US Wildfire Cases [1992 - 2015]

features (MONTH and DAY_OF_WEEK). We also subtract DISCOVERY_TIME from CONT_TIME to make a new feature DURATION. This attribute could be utilized to calculate the amount of wildfire damage.

As well as the feature selection and extraction, it is important to process missing values. We use PIPS_CODE to check the county in which the wildfire occurred. There are many missing values in the PIPS_CODE column. Fortunately, each wildfire case includes LATITUDE and LONGITUDE even when there is no state info. We find county using LATITUDE and LONGITUDE with the geopy library as shown in Figure 3.

After the feature selection, we check the features' correlation. Correlation Coefficient shows the relationship between features. High positive value means there is positively strong correlation, meanwhile high negative value means negatively strong correlation between features.

In the above matrix Figure 8, strong correlations are darker. We think that there will be very strong relationships between features, but the result is not the same as our expectation. There is an interesting links between month and latitude, probably this is because weather and season are related each other. Also, there is interesting correlation result between label and longitude. While there is negative correlation between longitude and month.

The range of the value of each feature varies greatly, and the magnitude of the absolute value can show a significant difference. Data normalizing gives features equivalent influence and this process will make the model find the hidden pattern easily. As shown in Figure 5-b, we normalize our data so that the size of the states will not influence the number of wildfire cases due to the fact that geographically larger states are more likely to have a higher number of wildfire cases.

3.2 Weather Dataset

We want to look deeper into our state wildfire cases, Colorado. The greatest portion of wildfire in Colorado comes from lightning as Figure 13. We entirely agree that we need other features to explain lightning occurrences. One of the biggest reason that comes out from the brainstorming is weather.Because the lightning tends to be occurred in the dry and humid weather condition.

We struggle to find weather datasets and related datasets from 1981-2010 Normals products from National Oceanic and Atmospheric Administration (NOAA). Typically, climate normals are defined as 30 years averages of meteorological conditions, such as air temperature, precipitation and etc. [1] [3] Among them, we selected temperature and precipitation datasets that have high possibility to impact wildfires occurrences' condition. Especially, we choose all three precipitation datasets, 25 percentile, 50 percentile and 75 percentile. We regard 25 and 75 percentile datasets more tend to explain special and severe condition that make occur wildfires. Followings are detail descriptions of each file:

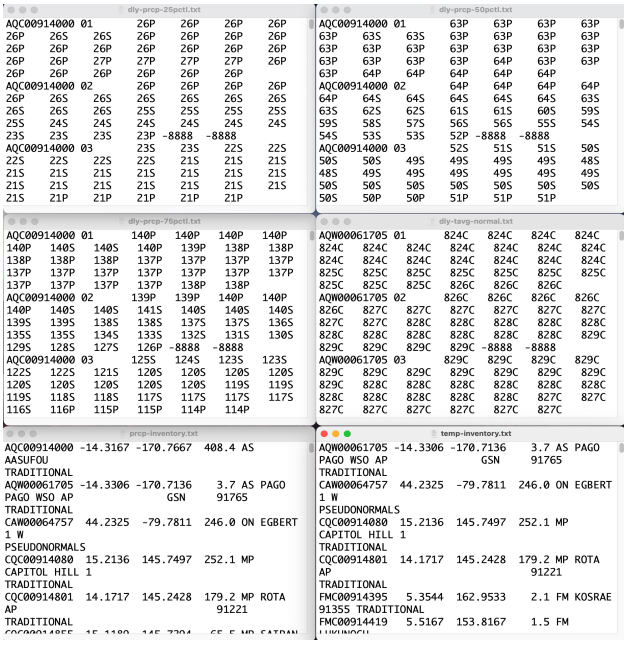


Figure 6: Weather Datasets

- **dly-prcp-25pctl**: Daily precipitation climatological 25th percentile data from 1981 to 2010. Contains each day normalized precipitation data for all stations.
- **dly-prcp-50pctl**: Daily precipitation climatological 50th percentile data from 1981 to 2010. Contains each day normalized precipitation data for all stations.
- **dly-prcp-75pctl**: Daily precipitation climatological 75th percentile data from 1981 to 2010. Contains each day normalized precipitation data for all stations.
- **dly-tavg-normal**: Daily mean temperature (average of temperature maximum and temperature minimum) climatological Average data from 1981 to 2010. Contains each day average temperature data for all stations.
- **prcp-inventory**: Contains all stations(9,307 stations) used in the precipitation analysis and its longitude, latitude, State and etc. information. But we just utilize the enumerated features.
- **temp-inventory**: Contains all stations(7,501 stations) used in the temperature analysis its longitude, latitude, State and etc information. But we just utilize the enumerated features.

```
[10]: dist = []
      loc_index = []

      for i in np.arange(len(df_weather_loc)):
          list1 = []
          for j in np.arange(len(df_inv_loc)):
              a = (df_inv_loc.loc[j, 'Latitude'], df_inv_loc.loc[j, 'Longitude'])
              b = (df_weather_loc.loc[i, 'LATITUDE'], df_weather_loc.loc[i, 'LONGITUDE'])
              list1.append(distance.distance(a,b).km)
          dist.append(min(list1))
          loc_index.append(list1.index(min(list1)))

      df_weather_loc['Distance'] = pd.DataFrame(dist)
      df_weather_loc['Inventory_Index'] = pd.DataFrame(loc_index)
      df_weather_loc
```

[10]:

	LATITUDE	LONGITUDE	Distance	Inventory_Index
0	39.292222	-105.183056	11.460354	30
1	38.913333	-105.983611	11.916755	4
2	39.100278	-106.367500	1.577794	196
3	37.345000	-102.805833	6.206311	184
4	38.891111	-105.431944	3.815104	113
...	...	...	...	...
34152	38.609900	-104.807800	24.524601	237
34153	39.421667	-103.878333	29.350907	235
34154	37.545900	-105.007100	20.897144	202
34155	40.394700	-104.564600	11.457235	86
34156	40.819800	-105.084700	19.735434	200

34157 rows × 4 columns

Figure 7: Calculating the distance between inventories and wildfires' spots

Regarding the precipitation percentile calculation, lower quartile (i.e., 25th percentile), median (50th percentile), and upper quartile (75th percentile) are calculated following standard procedures. If values from all 30 years are available, then the lower quartile is the eighth lowest value, the median is the average of the 15th and 16th lowest values, and the upper quartile is the 23rd lowest value.

The datasets are quite rough and raw, so we need to clean those to merge and utilize with wildfires dataset. With the weather datasets, first of all, we extract just Colorado state stations name from inventory files. And merge it to the precipitation and temperature datasets (weather datasets). Next, we extract the longitude, latitude information from wildfires dataset to find out the nearest stations

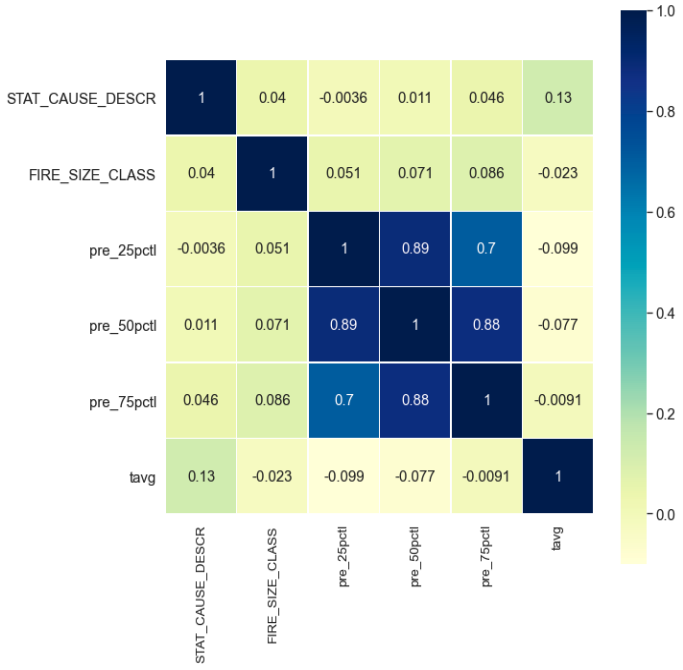


Figure 8: Selected Weather Features' Correlation Matrix

for each wildfire case by calculating with the distance method in geopy library. Lastly, with the nearest inventory information, we merge the weather datasets contains inventory name information to the wildfires dataset according to the exact wildfire discovery date.

We explore the new dataset and extract several features and make a new data frame to see how weather affects the wildfires occurrences and its size. For this reason, the new dataset contains cause of wildfires (i.e., STAT_CAUSE_DESCR), wildfires size classes (i.e., FIRE_SIZE_CLASS), 25th percentile precipitation, (i.e., pre_25pctl),

t-statistic / p-value	STAT_CAUSE_DESCR	FIRE_SIZE_CLASS
pre_25pctl	81.98 / 0.0	277.67 / 0.0
pre_50pctl	174. 93 / 0.0	419.79 / 0.0
pre_75pctl	391.49 / 0.0	520.86 / 0.0
tavg	698.78 / 0.0	776.10 / 0.0

Figure 9: T-statistic and Two-sided p-value

50th percentile precipitation (i.e., pre_50pctl), 75th percentile precipitation(i.e., pre_75pctl), and average temperature (i.e., tavg). We then check the correlation matrix between features. As to the result, the overall values are improved than previous result. Understandably, precipitation features are strongly correlated. Also, we find out that STAT_CAUSE_DESCR and tavg features are having 0.13 value of correlation. This value is not quite high, so we implement t-test for STAT_CAUSE_DESCR and FIRE_SIZE_CLASS with each weather features. The result of t-test are t-statistic and two-sided p-value. Each value of the result is that t-statistic is very high and two-sided p-value is 0. For this reaon, we can reject the null hypothesis that the means of STAT_CAUSE_DESCR, FIRE_SIZE_CLASS with each weather feature is the same and conclude that there is a statistically significant difference between the each models.

4 DATA MINING

This is the process of extracting valuable information by systematically and automatically analyzing statistical rules or patterns in data stored at a large scale. We establish the relationship between different attributes.

4.1 Classification

Classification is a kind of supervised learning, which is a process of grasping the category relationship of existing data and self-determining the category of newly observed data. We have multiple candidate models, such as Gradient Boosting, Random Forest, AdaBoost, Decision Tree, Multilayer Perceptron (MLP), K-Neighbors, and so on. The dataset provides FIRE_SIZE_CLASS having 7 classes, so we can test multi class classification. However, due to the small size of certain classes, we redefine the dataset into 4 classes as shown in Figure 10. We also utilize our created features (DAY_OF_WEEK and DAY_OF_MONTH).

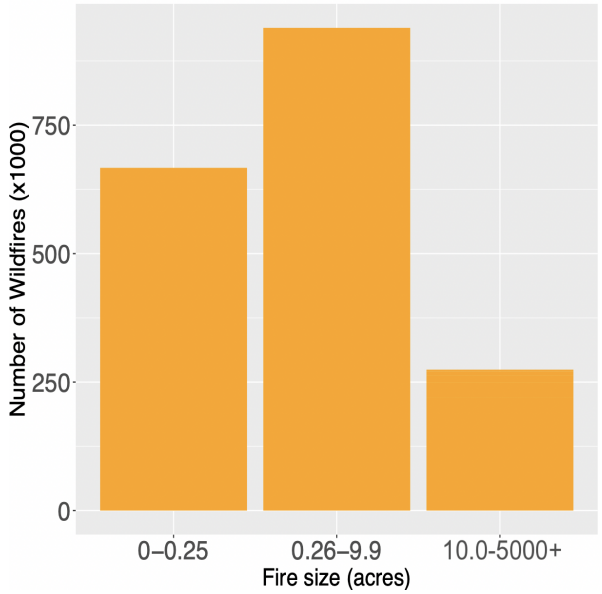


Figure 10: Classification by Size

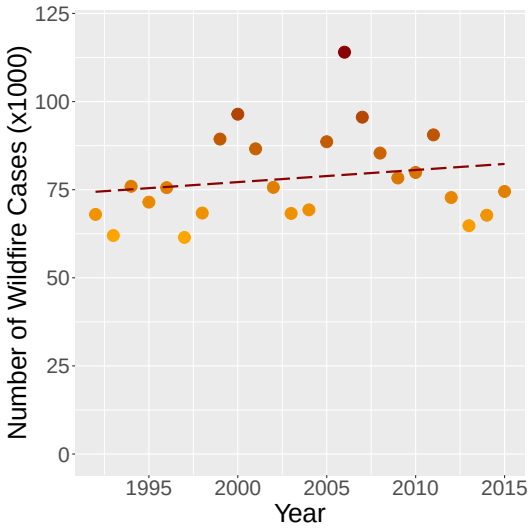


Figure 11: US Wildfire Cases

4.2 Clustering

Clustering refers to the process of dividing objects into several subgroups. In this project, we plan to use clustering in order to cluster states information based on the magnitude of damage the wildfire has caused in terms of acres.

One of the most simplest and powerful clustering method is k-Nearest-Neighbor. Here are the key factors of k-Nearest-Neighbor. In a data set, If an object has n features, the it regards as it is a point in a n-dimension space. Objects that have a similar tendency are prone to be close and it can be quantifiable with math formulations, such as Euclidean distance. Objects can be classified by k-nearest point around it. The result is dependable upon the parameter manipulation. Lastly, when we adjust the parameters and k value is set small, it could cause over-fitting. In contrary, if k value is set large, it could cause under-fitting.

In this project we can make several k-Nearest-Neighboring classifiers for Label feature or Fire size feature. With the trained classifiers we can predict other cases whether it is big fire or not, or cause of the fire (label).

4.3 Regression

Regression is based on the relevance of different features that help predict a continuous outcome. We can use regression to predict the damage scale of wildfires.

We check the correlation between selected features as Figure 4. There are several set of features show their slight correlation such as LONGITUDE and STAT-CAUSE-DESCR. We can train the data set and predict with trained regression model to find out whether our model works well or not. Figure 14 shows the

4.4 Visualization

Here we put our findings in a visual context for better understanding. As shown in Figure 11, there is a slight increase in the number of wildfire cases during the past two decades. Out of the

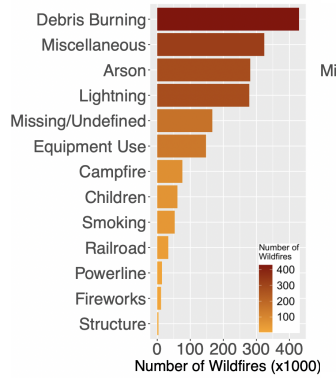


Figure 12

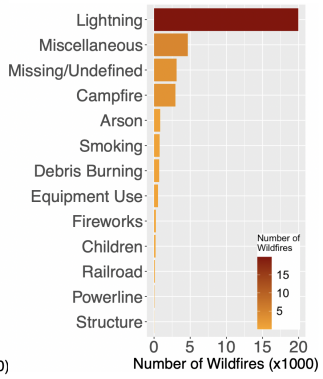


Figure 13

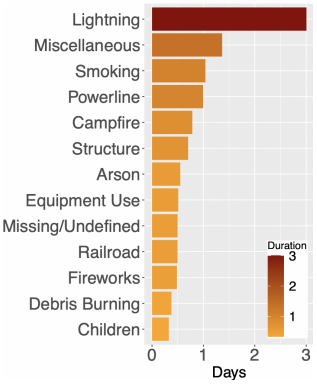


Figure 14

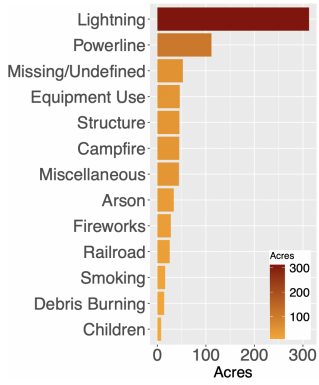


Figure 15

vast number of wildfire cases, Figure 12 shows that even though there is downward trend in the rate of arson wildfires on national forests [6] human caused wildfires still represent the majority of wildfire cases as mentioned in [4]

Figure 14 and Figure 15 shows the wildfires causes with respect to the wildfire duration and the wildfire causes with respect to the size of the wildfires respectively. We can tell that lighting and naturally caused fires tend to burn more than human caused fires. This can be explained by the responsiveness of how fast human can act to contain the fires.

Due to the recent pair of wildfires in the vicinity of and in Cameron Peak, both setting records as the second and largest wildfires respectively in the Colorado state history, in Figure 13 we provide an analysis on Colorado state wildfires. Lightening takes the most portion of cause in Colorado wildfire cases. This is interesting due to the fact that in Figure 12, we can see that human caused wildfires were dominant. Figure 16 shows the county view of the number of wildfires occurred in Colorado. We can see that most of the wildfires occur in the west where the mountains are located. This also connects to the result showing that lighting is responsible for most of the Colorado wildfires.

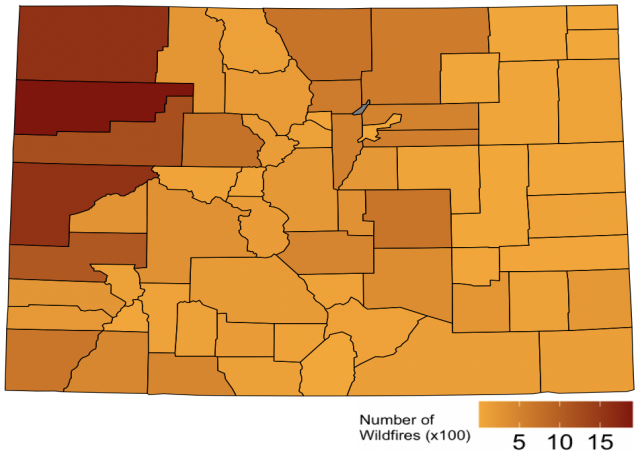


Figure 16: Colorado Wildfire Cases

5 EVALUATION

5.1 Tools

Here we list and describe the tools we utilize for this project. To interact with our data, we use Python and R with their large pool of libraries. Pandas, Scipy, SQLite3 provide Python to have the ability to easily manipulate and analyze our data flexibly. Our Python work environment is established in the Jupiter Notebook provided in this project.

After analyzing the data, we put our findings into visual context to quickly understand and evaluate what are meaningful feature or not by utilizing matplotlib and seaborn which are libraries for the seamless data visualization and plotting for Python . We also use libraries of R, such as RSQLite, dplyr to process our dataset and use ggplot2, maps, and mapdata for the visualization as well. You can run the R code (wildFireRun.R) we have provided via terminal. 1~2

5.2 Prediction

	Acre	Recall	Precision	Accuracy	F1-score
Gradient Boosting	0~0.25	0.68	0.59	0.72	0.63
	0.26~2	0.43	0.46	0.63	0.44
	2~5000+	0.49	0.54	0.72	0.51
Random Forest	0~0.25	0.68	0.60	0.73	0.64
	0.26~2	0.44	0.46	0.63	0.45
	2~5000+	0.48	0.54	0.72	0.51
AdaBoost	0~0.25	0.68	0.58	0.71	0.62
	0.26~2	0.38	0.43	0.62	0.41
	2~5000+	0.48	0.52	0.71	0.50
Decision Tree	0~0.25	0.58	0.58	0.70	0.58
	0.26~2	0.43	0.43	0.61	0.43
	2~5000+	0.49	0.49	0.69	0.49
Multi-layer Perceptron	0~0.25	0.70	0.57	0.70	0.63
	0.26~2	0.41	0.45	0.63	0.43
	2~5000+	0.45	0.53	0.72	0.49
K-Neighbors	0~0.25	0.70	0.57	0.71	0.63
	0.26~2	0.43	0.45	0.62	0.44
	2~5000+	0.44	0.56	0.72	0.49

Among the various factors in the dataset, we want to predict the fire size class to approximate the scale of damage caused by wildfires. As related factors, latitude, longitude, month, day of week, state, and cause were used. As can be seen from the difference in the number of fire size classes, data is concentrated in 0.26 9.9 acres, accounting for 50% of all cases. If the data is oriented to one side like this, an error occurs that biases the prediction during training. Therefore, it is possible to reduce the difference in the number of data for each class by dividing the fire size class into 3 largely, small damage (0 0.25), medium damage (0.26 2), and large damage (2 5000+). Although the number of wildfire occurrence cases was large, the prediction results of each model were insufficient due to the lack of appropriate factors for predicting the size of the forest fire and the small difference between fire size classes.

6 CONCLUSION

In this project, we use the 1.88 Million Wildfire dataset to analyze and provide a prediction model using machine learning and deep learning techniques that help us estimate the scale of damage caused by wildfires. Individual fires are partially randomly distributed in space and time, variation exists over small and large spatial scales and across short and long temporal scales. Development of statistical models that can predict such variations is needed however, is a more challenging endeavor that can nevertheless yield some advances in our understanding of wildfires.

7 APPENDIX

Work Done By Individual Group Members:

Most of the work was done together and each individuals have equally contributed to this project, where the major tasks of each individual are as follows:

Chanheum Park: Data visualization using R.

Feelgyun Oh: Data analysis using Python and introducing the weather dataset for an in depth analysis of Colorado state from our dataset.

Taeho Kim: Data preprocessing and providing classification models using machine learning and deep learning techniques using Python.

REFERENCES

[1] Anthony Arguez, Scott Applequist, RS Vose, Imke Durre, MF Squires, and Xungang Yin. 2012. NOAA's 1981–2010 climate normals: methodology of temperature-related normals. *NCDC Report* 7 (2012).

[2] Jennifer K Balch, Bethany A Bradley, John T Abatzoglou, R Chelsea Nagy, Emily J Fusco, and Adam L Mahood. 2017. Human-started wildfires expand the fire niche across the United States. *Proceedings of the National Academy of Sciences* 114, 11 (2017), 2946–2951.

[3] I Durre, MF Squires, RS Vose, A Arguez, S Applequist, and X Yin. 2013. Computational procedures for the 1981–2010 normals: precipitation, snowfall, and snow depth. *National Climatic Data Center Report (11 pp.)* (2013).

[4] Jon E Keeley and Alexandra D Syphard. 2018. Historical patterns of wildfire ignition sources in California ecosystems. *International journal of wildland fire* 27, 12 (2018), 781–799.

[5] Claire Miller, Matt Plucinski, Andrew Sullivan, Alec Stephenson, Carolyn Huston, Kay Charman, Mahesh Prakash, and Simon Dunstall. 2017. Electrically caused wildfires in Victoria, Australia are over-represented when fire danger is elevated. *Landscape and Urban Planning* 167 (2017), 267–274.

[6] Jeffrey P Prestemon, Todd J Hawbaker, Michael Bowden, John Carpenter, Maureen T Brooks, Karen L Abt, Ronda Sutphen, and Samuel Scranton. 2013. Wildfire ignitions: a review of the science and recommendations for empirical modeling. *Gen. Tech. Rep. SRS-GTR-171. Asheville, NC: USDA-Forest Service, Southern Research Station. 20 p.* 171 (2013), 1–20.

[7] Alexandra D Syphard and Jon E Keeley. 2015. Location, timing and extent of wildfire vary by cause of ignition. *International Journal of Wildland Fire* 24, 1 (2015), 37–47.

[8] Rachael Tatman. 2020. 1.88 Million US Wildfires. https://www.kaggle.com/rtatman/188-million-us-wildfires