Introduction

Wildfires have broken out all over the western region of the United States in 2020, devastating communities and creating smoke plumes that can be seen even from space via satellite images. Some billows of smoke have carried over to places as far away as London, and it has been said that cities such as San Francisco and Seattle have some of the lowest quality of air on the entire planet currently due to the fires. As one of us lives close to the Bobcat Fire in California, which has currently burned over 93 thousand acres of land, the team thought it would be an interesting topic to delve into the data that has been collected in this domain over the past quarter century or so to mine any insights.

US Wildfires Data Set

PLACEHOLDER DESCRIPTION FROM KAGGLE COMPETION

"*This data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015. It is the third update of a publication originally generated to support the national Fire Program Analysis (FPA) system. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. The following core data elements were required for records to be included in this data publication: discovery date, final fire size, and a point location at least as precise as Public Land Survey System (PLSS) section (1-square mile grid). The data were transformed to conform, when possible, to the data standards of the National Wildfire Coordinating Group (NWCG). Basic error-checking was performed and redundant records were identified and removed, to the degree possible. The resulting product, referred to as the Fire Program Analysis fire-occurrence database (FPA FOD), includes 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned during the 24-year period.*"

(https://www.kaggle.com/rtatman/188-million-us-wildfires (https://www.kaggle.com/rtatman/188-million-us-wildfires))

Objective One

Question of Interest

Given the size, location, date and other relevant features from the dataset, can we predict the cause of a wildfire?

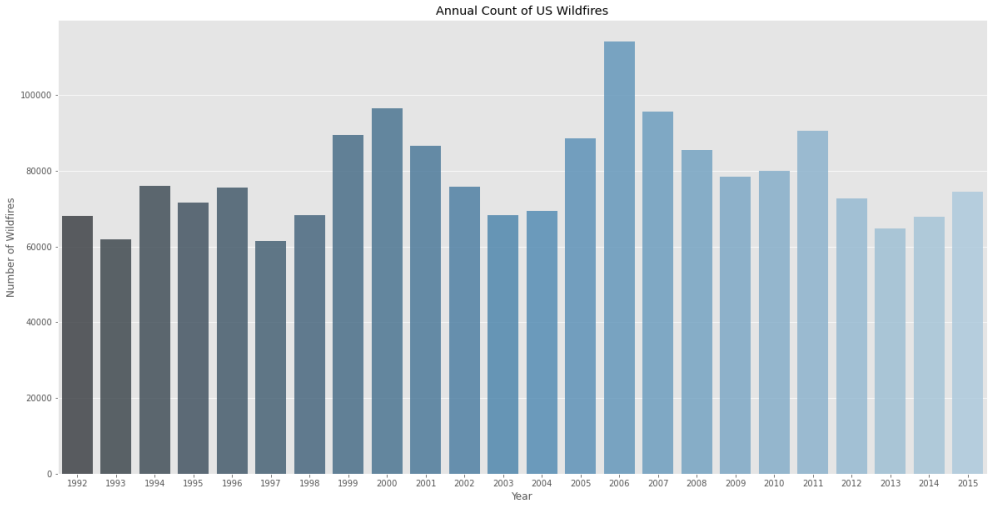
Exploratory Data Analysis

Total US Wildfires by Year

The dataset includes wildfires that occurred in the United States from **1992** to **2015**.

In [25]:
```
from IPython.display import Image
Image(filename='./img/us_wilfires_by_year.png')
```

Out[25]:

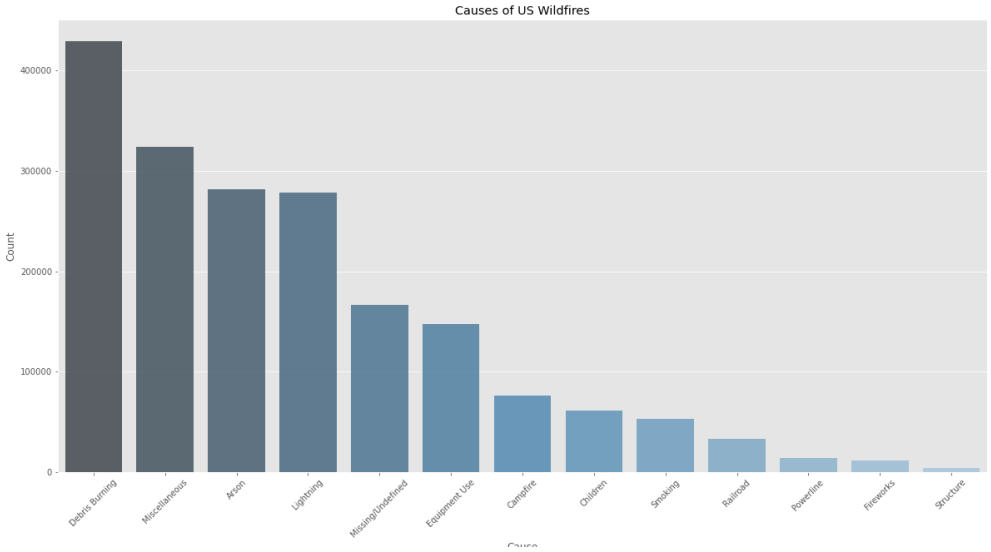


Total US Wildfires by Cause

Across all of the states included in the dataset, **debris burning** was the category identified as having caused the most wildfires. The lables are obviously very skewed, so this may need to be taken into account when building our final prediction model. A technique like **SMOTE** might be of use so that the categories with the lower samples are artificially upsampled to match that of the hightest one.

In [26]:
```
Image(filename='./img/us_wilfire_causes.png')
```

Out[26]:

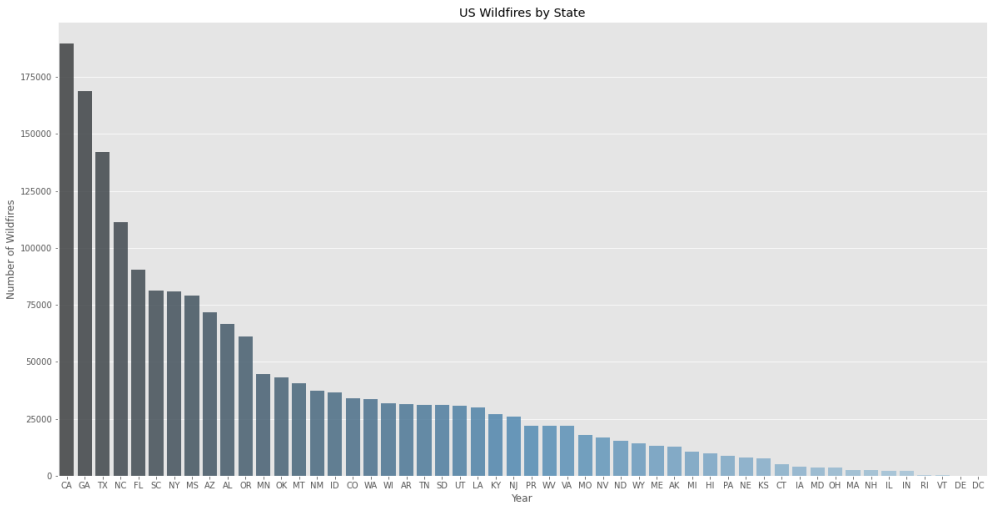


Total US Wildfires by State

California, Georgia and Texas have the highest volume of recorded wildfires. California has the highest population of any state in the union, followed by a distant second with Texas. The counts for those states could be somewhat expected in that they have higher populations. The counts for Georgia however do seem a bit high considering that their state population is lower on the list of most populated states. Further investigation might be warranted there. We may also look to see if states differ in their category rankings.

In [27]:
```
Image(filename='./img/us_wilfires_by_state.png')
```

Out[27]:



Modeling

Our current model has only been trained on data from California, we'll expand this to include data from all states and compare and contrast different algorithims and techniques. (such as the upsampling strategy mentiond above)

For example, the current build is using a Gaussian Naive Bayes classifer, we'll compare this to other models such Random Forrest. (which may or may not offer better predictive accuracy, but likely will come at the expense of processing time)

Predictions

Once the model has been defined, predictions can be made on new and unseen data. `Serverless Framework` to be discussed more in subsequent sections, which allows us to test a model locally as if it were in the cloud.

In [28]:
```
Image(filename='./img/sls_predictions.png')
```

Out[28]:

Objective Two

Now that a model has been defined, can we make use of cloud resources to make predictions?

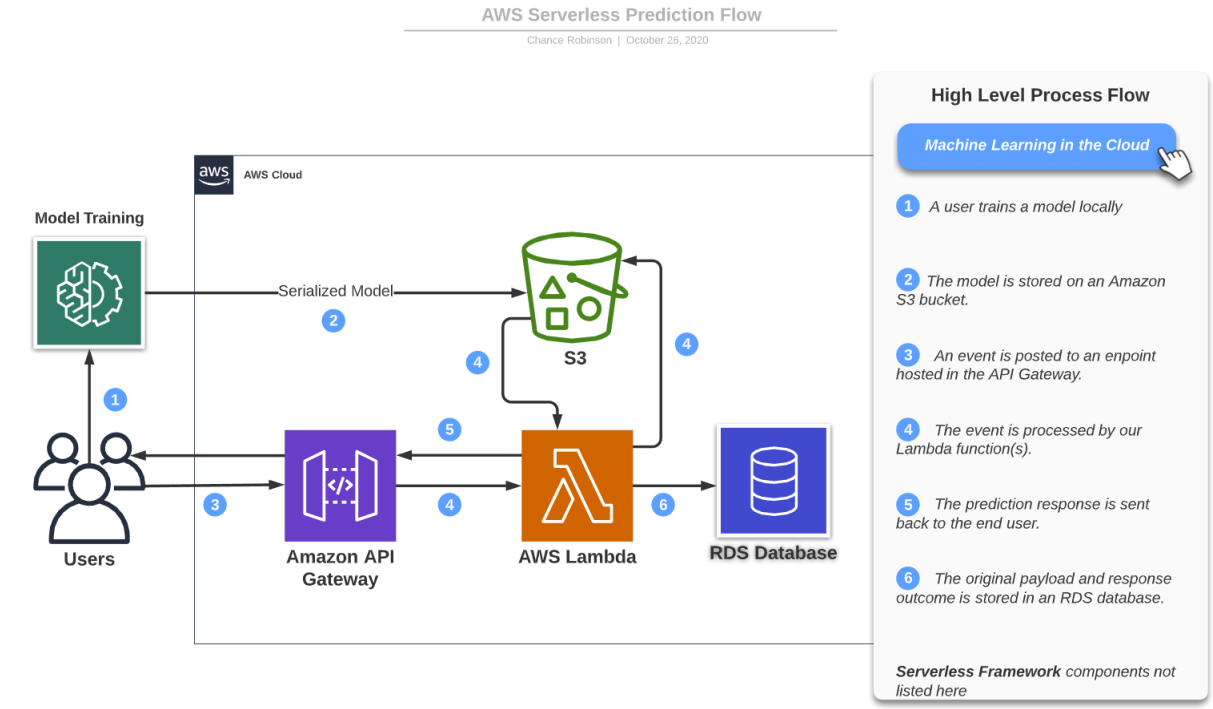
Cloud Deployment

- Serverless Framework
  - Local Setup and Validation
    - YAML (Blueprint)
    - Dependencies
    - ZIP File Size Reduction
    - Stages (DEV, QA, PROD deployment)
  - API Gateway
  - Lambda
  - RDS

AWS Serverless Process Flow

In [29]:
```
Image(filename='./img/aws_serverless_prediction_flow.png')
```

Out[29]:

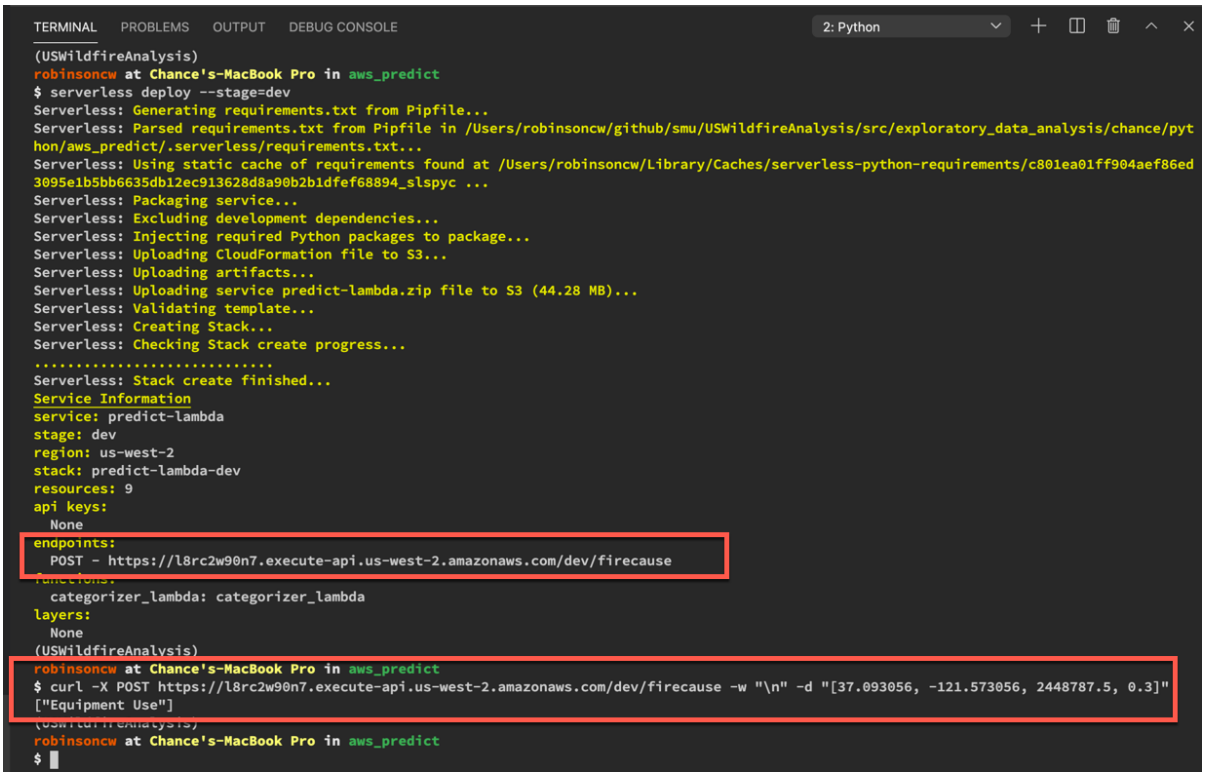


Serverless Deployment

Once deployed, a working URL endpoint will be retured which can be used to validate the solution via a `CURL` or other web service request.

In [30]:
```
Image(filename='./img/aws_prediction_output.png')
```

Out[30]:

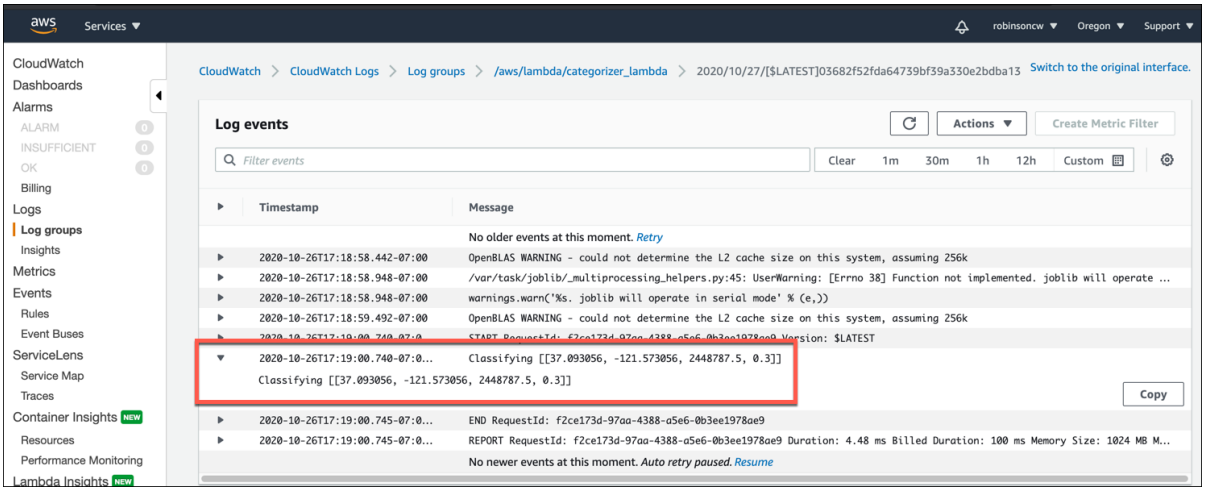


CloudWatch Logging

Logging can be reviewedthrough the CloudWatch area from the AWS console.

In [31]:
```
Image(filename='./img/aws_cloudwatch_logs.png')
```

Out[31]:



Appendix

Data Dictionary

This dataset is an SQLite database that contains the following information:

The source file can be found on Kaggle | 1.88 Million US Wildfires (https://www.kaggle.com/rtatman/188-million-us-wildfires).

Column Name	Data Type	Description
FOD_ID	Integer	Global unique identifier.
FPA_ID	String	Unique identifier that contains information necessary to track back to the original record in the source dataset.
SOURCESYSTEMTYPE	String	Type of source database or system that the record was drawn from (federal, nonfederal, or interagency).
SOURCESYSTEM	String	Name of or other identifier for source database or system that the record was drawn from. See Table 1 in Short (2014), or \Supplements\FPAFODsourcelist.pdf, for a list of sources and their identifier.
NWCGREPORTINGAGENCY	String	Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).
NWCGREPORTINGUNIT_ID	String	Active NWCG Unit Identifier for the unit preparing the fire report.
NWCGREPORTINGUNIT_NAME	String	Active NWCG Unit Name for the unit preparing the fire report.
SOURCEREPORTINGUNIT	String	Code for the agency unit preparing the fire report, based on code/name in the source dataset.
SOURCEREPORTINGUNIT_NAME	String	Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.
LOCALFIREREPORT_ID	String	Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.
LOCALINCIDENTID	String	Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.
FIRE_CODE	String	Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/ (https://www.firecode.gov/)).
FIRE_NAME	String	Name of the incident, from the fire report (primary) or ICS-209 report (secondary).
ICS209INCIDENT_NUMBER	String	Incident (event) identifier, from the ICS-209 report.
ICS209NAME	String	Name of the incident, from the ICS-209 report.
MTBS_ID	String	Incident identifier, from the MTBS perimeter dataset.
MTBSFIRENAME	String	Name of the incident, from the MTBS perimeter dataset.
COMPLEX_NAME	String	Name of the complex under which the fire was ultimately managed, when discernible.
FIRE_YEAR	Integer	Calendar year in which the fire was discovered or confirmed to exist.
DISCOVERY_DATE	Float	Date on which the fire was discovered or confirmed to exist.
DISCOVERY_DOY	Integer	Day of year on which the fire was discovered or confirmed to exist.
DISCOVERY_TIME	String	Time of day that the fire was discovered or confirmed to exist.
STATCAUSECODE	Float	Code for the (statistical) cause of the fire.
STATCAUSEDESCR	String	Description of the (statistical) cause of the fire.

Column Name	Data Type	Description
CONT_DATE	Float	Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).
CONT_DOY	Float	Day of year on which the fire was declared contained or otherwise controlled.
CONT_TIME	String	Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).
FIRE_SIZE	Float	Estimate of acres within the final perimeter of the fire.
FIRESIZECLASS	String	Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
LATITUDE	Float	Latitude (NAD83) for point location of the fire (decimal degrees).
LONGITUDE	Float	Longitude (NAD83) for point location of the fire (decimal degrees).
OWNER_CODE	Float	Code for primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
OWNER_DESCR	String	Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
STATE	String	Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.
COUNTY	String	County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.
FIPS_CODE	String	Three-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities.
FIPS_NAME	String	County name from the FIPS publication 6-4 for representation of counties and

References

Kaggle | 1.88 Million US Wildfires (https://www.kaggle.com/rtatman/188-million-us-wildfires)