Wildfires

A capstone project by:

Marty Hauck, John Osire, Mark Harris, Harold Haugen, Nahid Boustani, Osie David

Georgetown University Certificate in Data Science

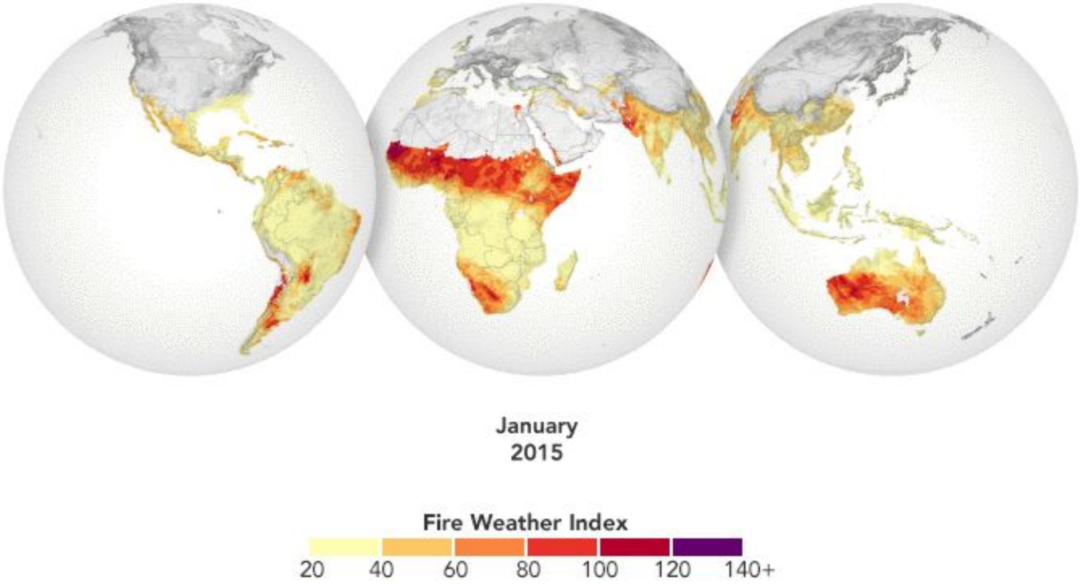
June 20, 2020

In the “OtherSpace” dimension

Like a prototype of organizational partnership, this endeavor made us further acknowledge the degree of willingness, effort, and integrity it takes to be both anthropocentric and data savvy; to hold the space for one another’s vision to manifest and to utilize our data to tell our story, the story of relentless traveling back and forth between our intuition and the spatial space, our vision and the scientific tools, our challenges and our accomplishments.

NB

A Hotter Drier World



Carbon Dioxide | Greenhouse Gases | Global Warming |

Health Risks from Smoke | Structural Destruction

Motivation

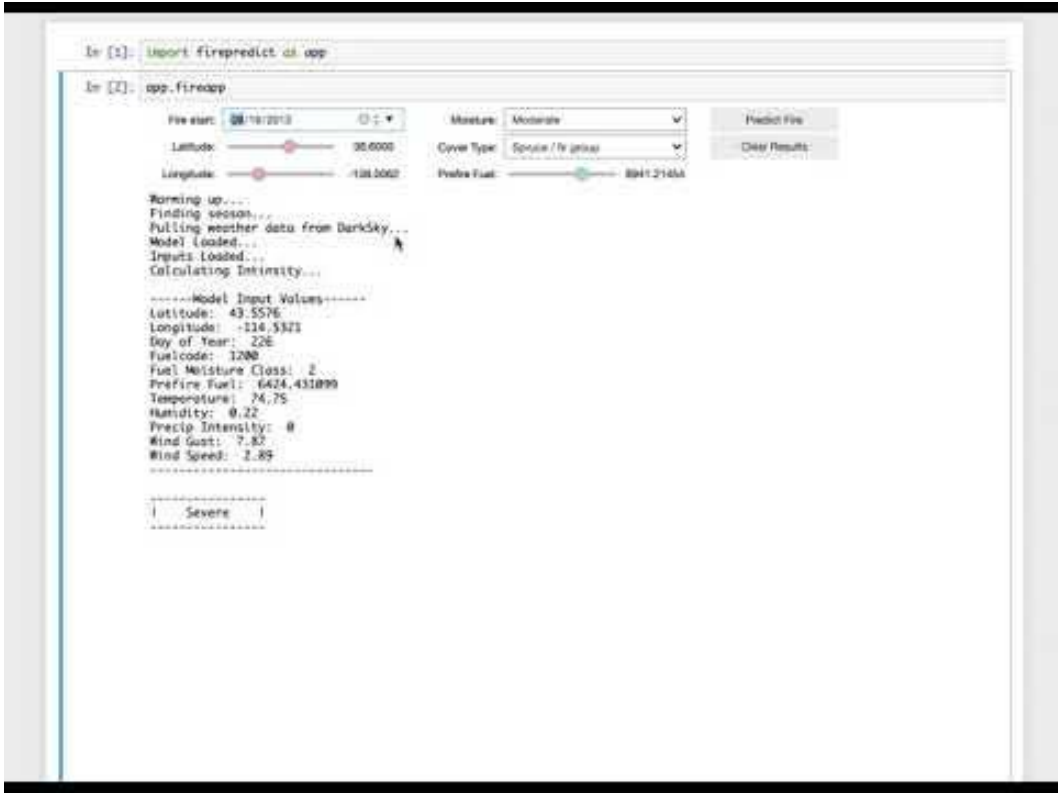
- Provide a new mean to the public to anticipate and respond to the dangers of fire
- Address the effects of this natural phenomenon in a novel way
  - Land observed fire severity data is collected post fire
  - NASA Earth Data provides Near Real Time fire observation
- Can we simulate Near Real Time fire intensity using land data?

Hypothesis & Application

- The goal of this project was to develop a predictive model to approximate the intensity of fire and create an easy and accessible tool to check Forest Fire Danger Index in Near Real Time across the United States.

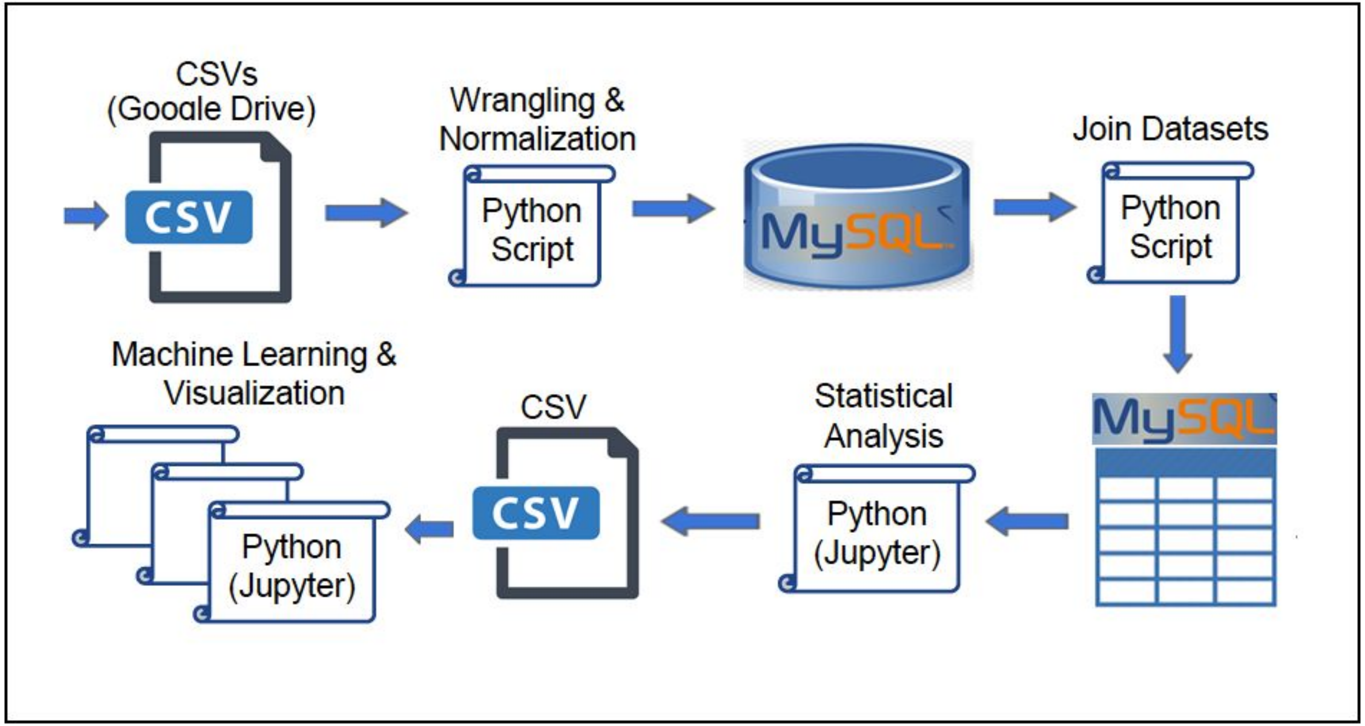
- Hypothesis: Forest Fire Danger Index will be influenced by
  - Land Coverage
  - Wind
  - Humidity
  - Temperature
- Application:
  - Tell us about where you're located and a bit about your land coverage and we'll predict the intensity of any possible fire in the days coming.

Fire Danger Forecaster

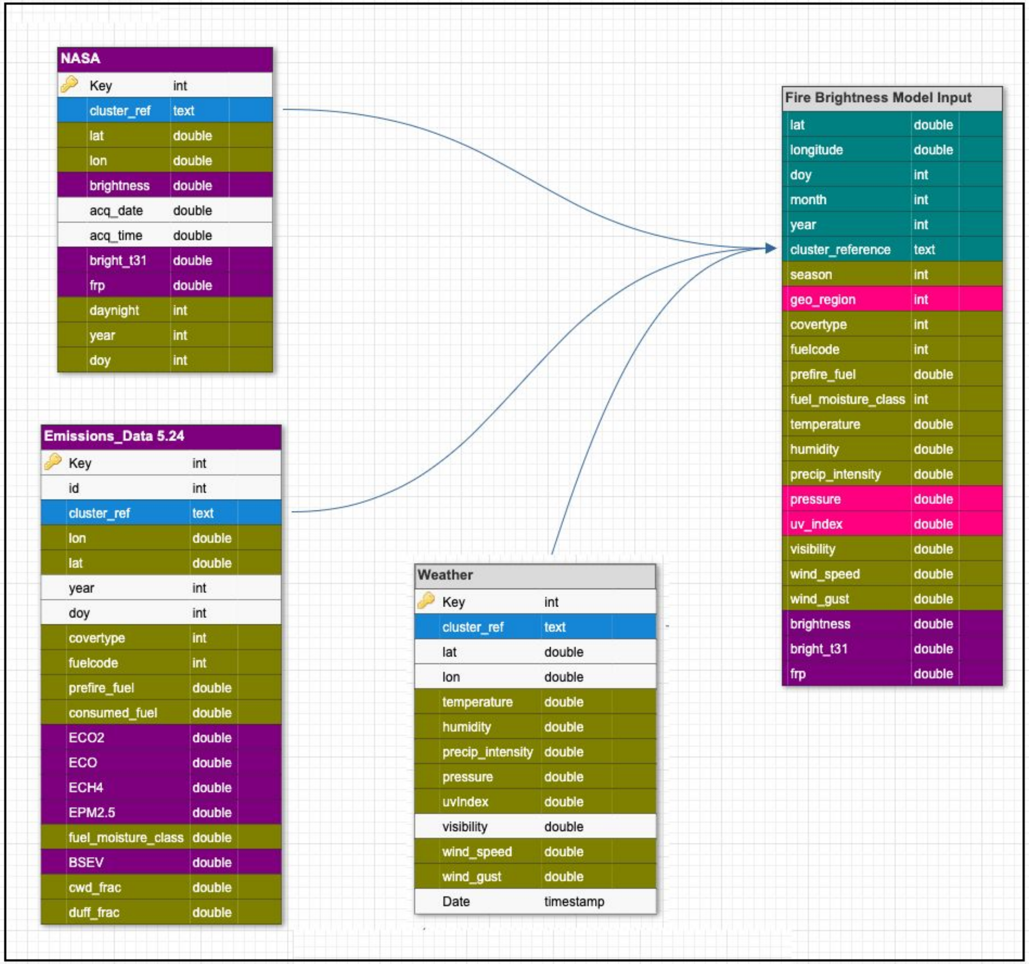


Data Storage

Data Storage



Data Model



NASA
Key int
cluster_ref text
lat double
lon double
brightness double
acq_date double
acq_time double
bright_t31 double
frp double
daynight int
year int
doy int
Emissions_Data 5.24
Key int
id int
cluster_ref text
lon double
lat double
year int
doy int
covertype int
fuelcode int
prefire_fuel double
consumed_fuel double
ECO2 double
ECO double
ECH4 double
EPM2.5 double
fuel_moisture_class double
BSEV double
cwd_frac double
duff_frac double
Weather
Key int
cluster_ref text
lat double
lon double
temperature double
humidity double
precip_intensity double
pressure double
uvindex double
visibility double
wind_speed double
wind_gust double
Date timestamp
Fire Brightness Model Input
lat double
longitude double
doy int
month int
year int
cluster_reference text
season int
geo_region int
covertype int
fuelcode int
prefire_fuel double
fuel_moisture_class int
temperature double
humidity double
precip_intensity double
pressure double
uv_index double
visibility double
wind_speed double
wind_gust double
brightness double
bright_t31 double
frp double

Ingestion & Wrangling

Ingestion & Wrangling

3 Primary Data Sources Used

- Emissions - *USDA Missoula Fire Lab Emission Inventory (MFLEI) for CONUS*
  (7.3m over 2003 - 2015)
- NASA - *MODIS/Aqua+Terra Thermal Anomalies/Fire locations*
  (2.15m over 2003 - 2019)
- Weather - *Dark Sky API*

Delivery Type

- CSV format downloaded from USDA/NASA website
- Developed Python script to interact with DarkSky API to pull historical weather information across lat/long/day/year

Tools

- DarkSky API Function, DBScan, K-D Tree, KMeans, Custom Functions

Ingestion & Wrangling

DBScan to Determine Centerpoint for Fire Events

- Use of unsupervised clustering to reduce 7.61m individual coordinate based records to 109k centerpoints *(assignment of unique cluster label, reduced weather API pull, assessment of fire event growth/size/description)*

Ref 1

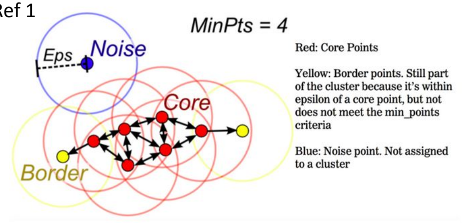
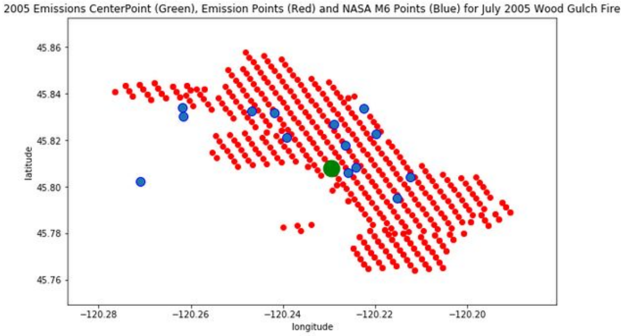
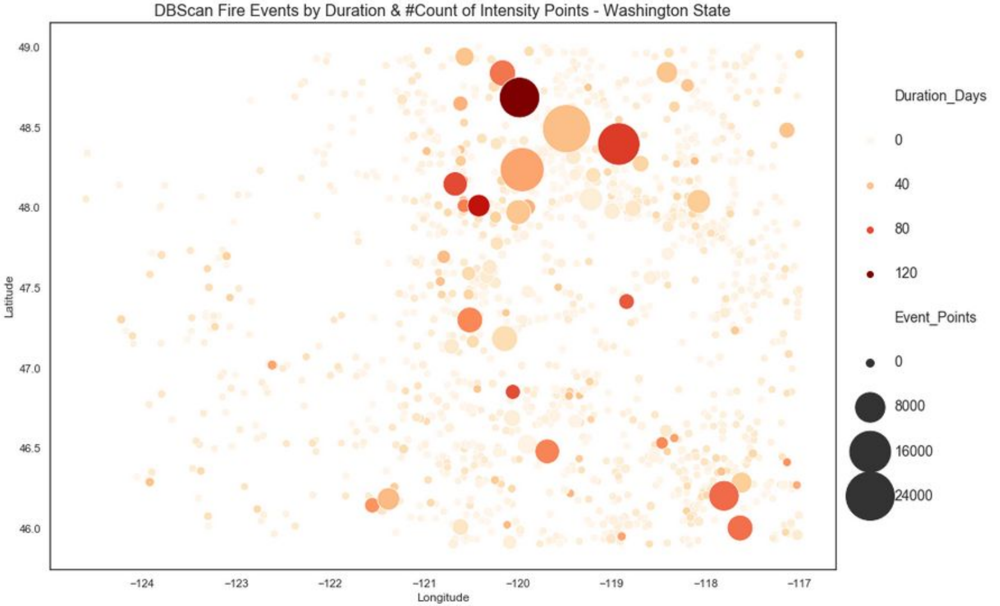


Figure 1 demonstrating density-based clustering



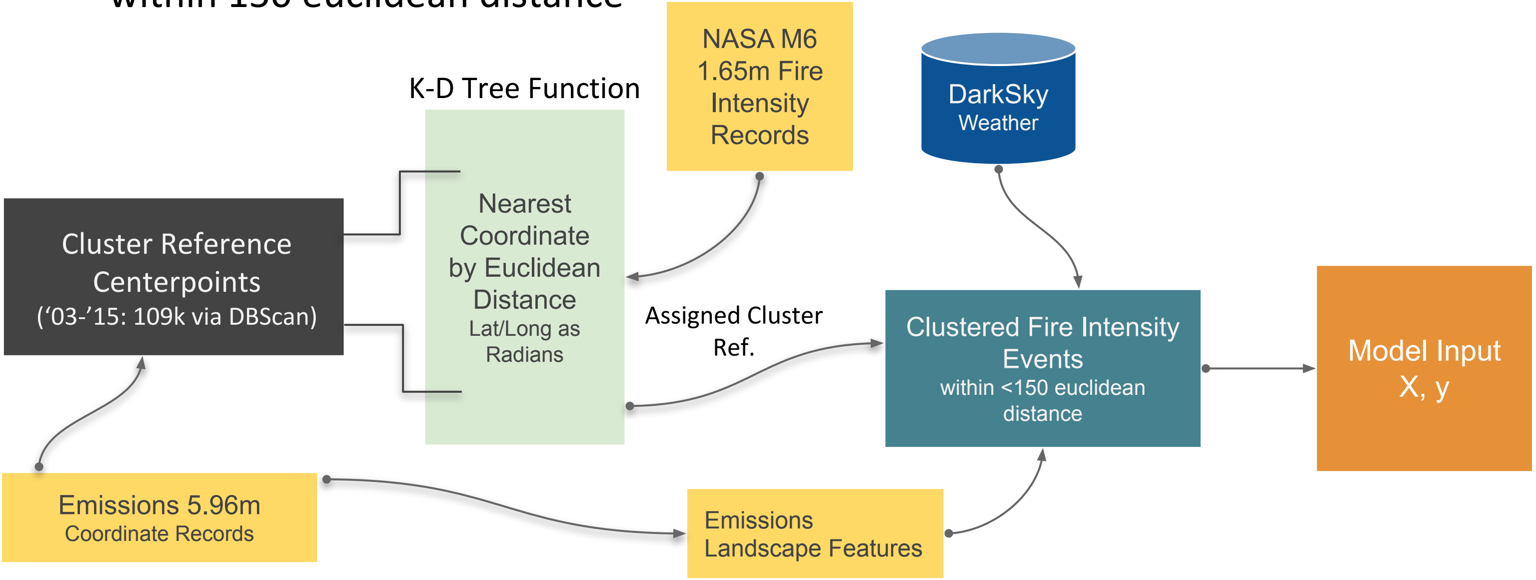


Ref 1 - https://medium.com/@agarwalvibhor84/lets-cluster-data-points-using-dbscan-278c5459bee5

Ingestion & Wrangling

K-D Tree to determine related NASA M6 records

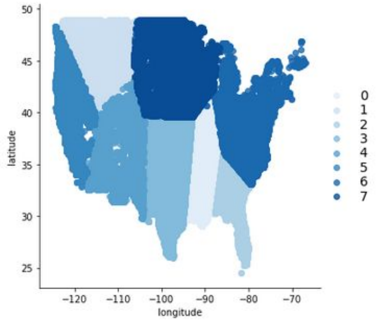
- Iterative process to determine closest cluster centerpoints for each NASA M6 record to Emissions/landscape cluster centerpoints
- Assignment of cluster reference label, distance and centerpoint lat/long for those within 150 euclidean distance



Ingestion & Wrangling

Other Wrangling Actions

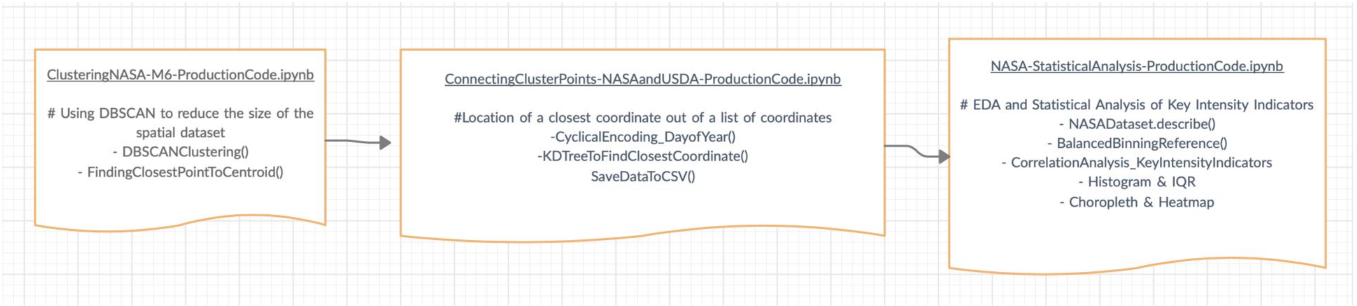
- KMeans to derive geographic region as optional feature
- Null Values, Merging, DeDuplication, Grouping, Selection by Max, Count
- Development of optional Y categories over various binning scenarios
- Date/Time manipulation, derivation of Season as optional feature; cyclical encoding of DOY
- Cluster review and optimization (consolidating low volume/ overlapping coordinates, DOY distribution e.g., high range vs low count)

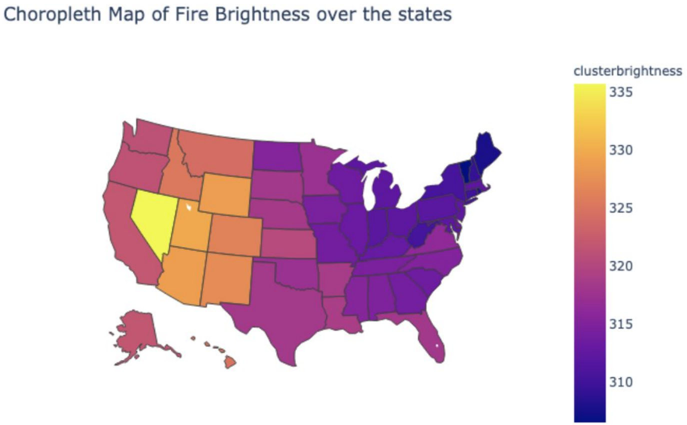


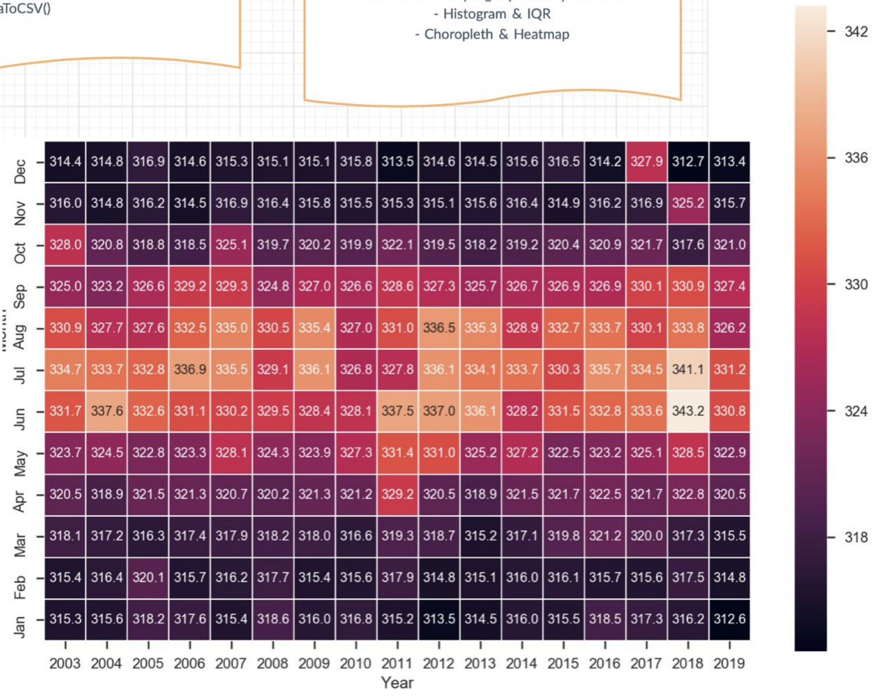
Ref 1 - https://medium.com/@agarwalvibhor84/lets-cluster-data-points-using-dbscan-278c5459bee5

Exploratory and Statistical Analysis

Geographical and Seasonal Changes of Fire Intensity



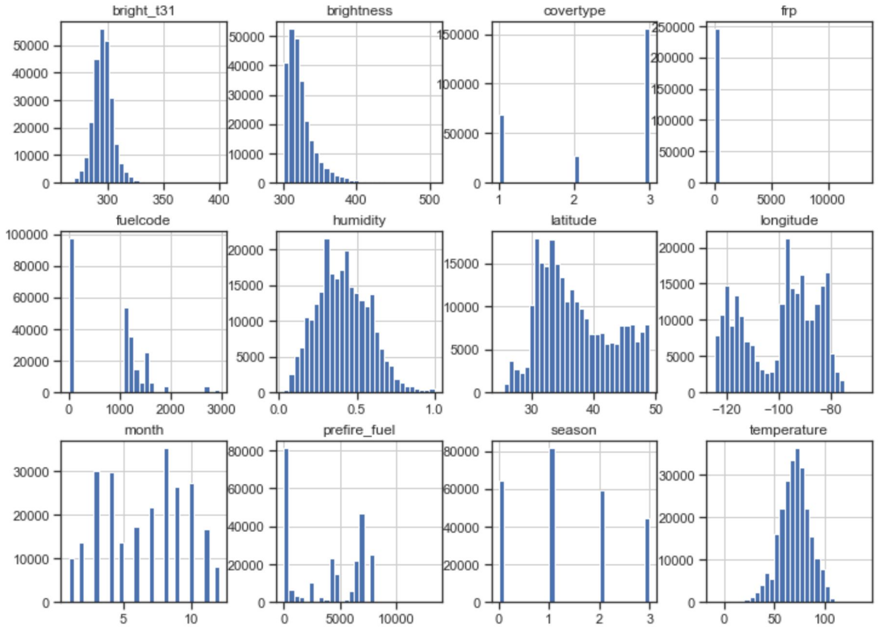




Month	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Dec	314.4	314.8	316.9	314.6	315.3	315.1	315.1	315.8	313.5	314.6	314.5	315.6	316.5	314.2	327.9	312.7	313.4
Nov	316.0	314.8	316.2	314.5	316.9	316.4	315.8	315.5	315.3	315.1	315.6	316.4	314.9	316.2	316.9	325.2	315.7
Oct	328.0	320.8	318.8	318.5	325.1	319.7	320.2	319.9	322.1	319.5	318.2	319.2	320.4	320.9	321.7	317.6	321.0
Sep	325.0	323.2	326.6	329.2	329.3	324.8	327.0	326.6	328.6	327.3	325.7	326.7	326.9	326.9	330.1	330.9	327.4
Aug	330.9	327.7	327.6	332.5	335.0	330.5	335.4	327.0	331.0	336.5	335.3	328.9	332.7	333.7	330.1	333.8	326.2
Jul	334.7	333.7	332.8	336.9	335.5	329.1	336.1	326.8	327.8	336.1	334.1	333.7	330.3	335.7	334.5	341.1	331.2
Jun	331.7	337.6	332.6	331.1	330.2	329.5	328.4	328.1	337.5	337.0	336.1	328.2	331.5	332.8	333.6	343.2	330.8
May	323.7	324.5	322.8	323.3	328.1	324.3	323.9	327.3	331.4	331.0	325.2	327.2	322.5	323.2	325.1	328.5	322.9
Apr	320.5	318.9	321.5	321.3	320.7	320.2	321.3	321.2	329.2	320.5	318.9	321.5	321.7	322.5	321.7	322.8	320.5
Mar	318.1	317.2	316.3	317.4	317.9	318.2	318.0	316.6	319.3	318.7	315.2	317.1	319.8	321.2	320.0	317.3	315.5
Feb	315.4	316.4	320.1	315.7	316.2	317.7	315.4	315.6	317.9	314.8	315.1	316.0	316.1	315.7	315.6	317.5	314.8
Jan	315.3	315.6	318.2	317.6	315.4	318.6	316.0	316.8	315.2	313.5	314.5	316.0	315.5	318.5	317.3	316.2	312.6

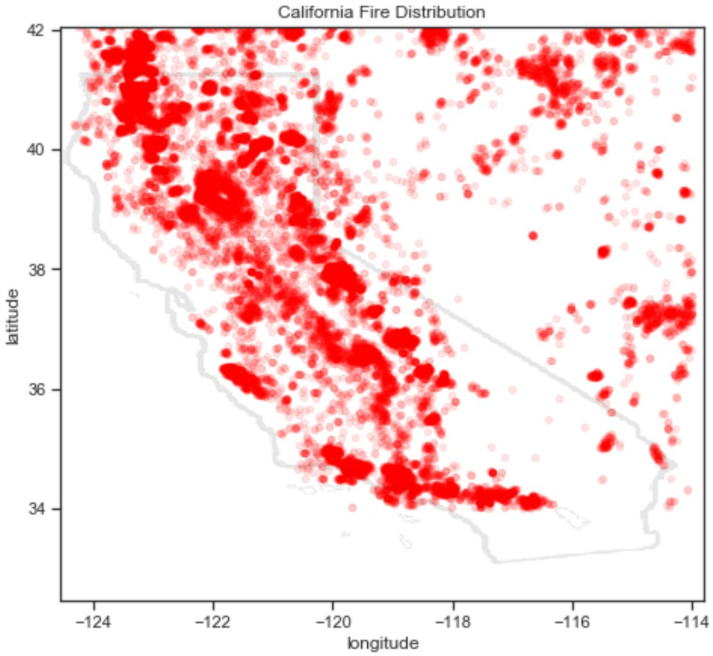
Exploratory Data Analysis

- Applied a consistent approach to categorical and numeric data by using Python visualization tools
  - Visualizing fire intensity dataframes with Pandas



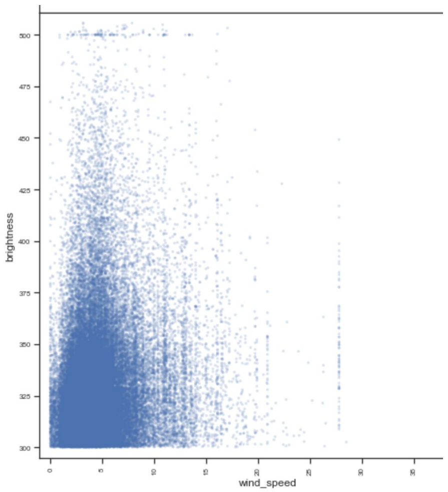
Exploratory Data Analysis

- Used Matplotlib wrappers to create visual depictions of fire event data

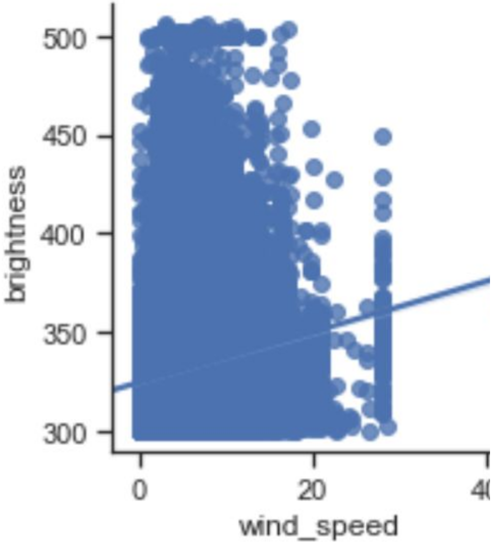


Exploratory Data Analysis

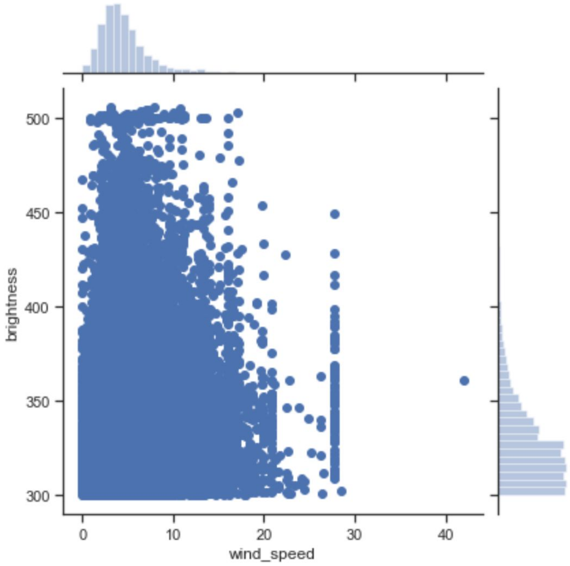
- Used statistical tools to illustrate the brightness feature is loosely concentrated around the median and is a function of wind speed



Scatter Matrix



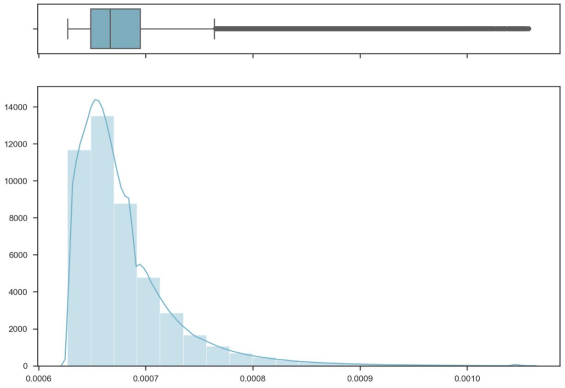
Pair Plot

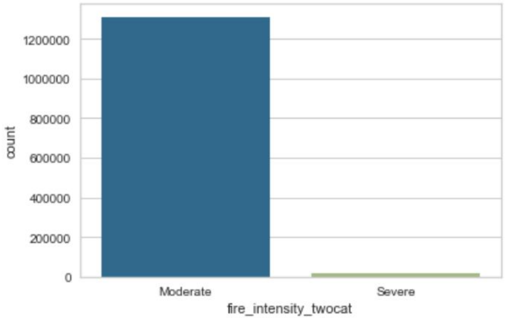


Joint Plot

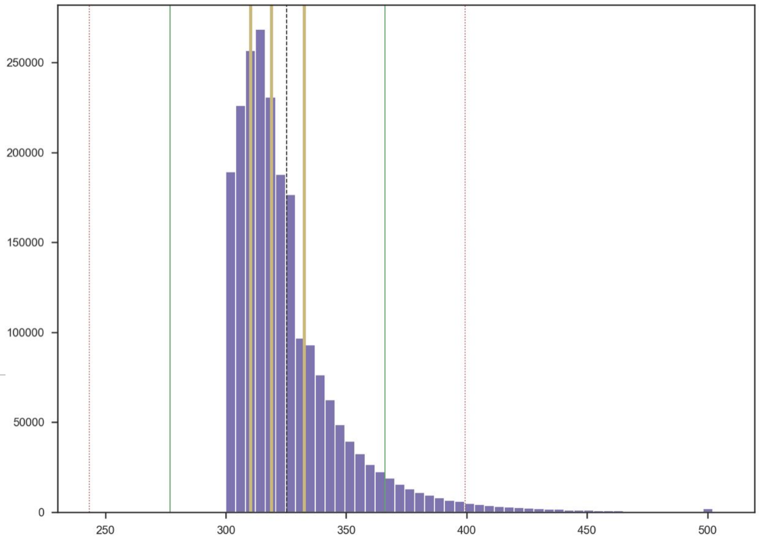
R2 Correlation of windspeed and brightness: 0.018661473

Binning the Y Variable





IQR Ranges: green and red lines
Mean: Dotted black line line
Binning Reference Values are: 310.30 ,318.90 , and 332.60 with 310.30 as Median value.



Feature Selection Modeling & Tuning

Feature Selection, Modeling & Tuning

From EDA:

- 3 Data Sources
- 14, 22 & 21 features

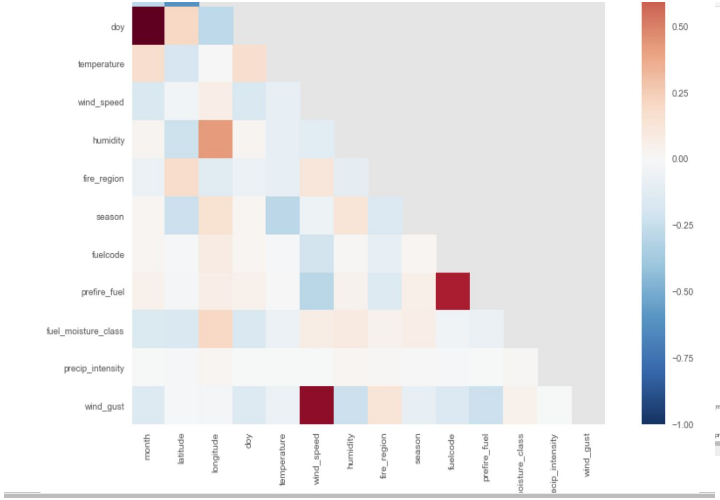
Research & Analysis

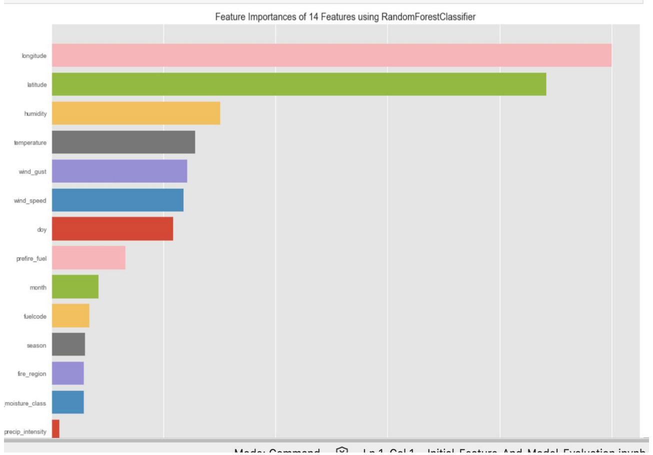
- Duplicative (time, time stamp)
- No value (weather icon)
- Not available at fire start (burned area)
- Not enough data (Precipitation Amount)
- Added Region and Season

- **Selected approx 15 features for further analysis [Location, Landscape, Time, Weather]**

Feature Selection, Modeling & Tuning

- Let’s explore and experiment -
  - Is predictive?
  - Highly correlated?
  - Needs scaling or encoding?
  - Favorite Graphs Pearson 2D Ranking & Feature Importance
  - **12 ish Features** (some surprises lat/lon over region)

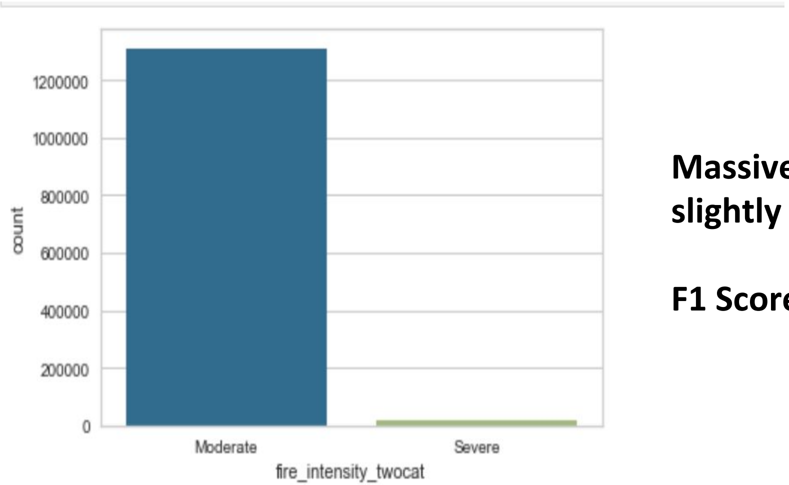




Feature Selection, **Modeling** and Tuning

- Classification: 2, 3 or 4 categories?

- 1.3 Millions Instances

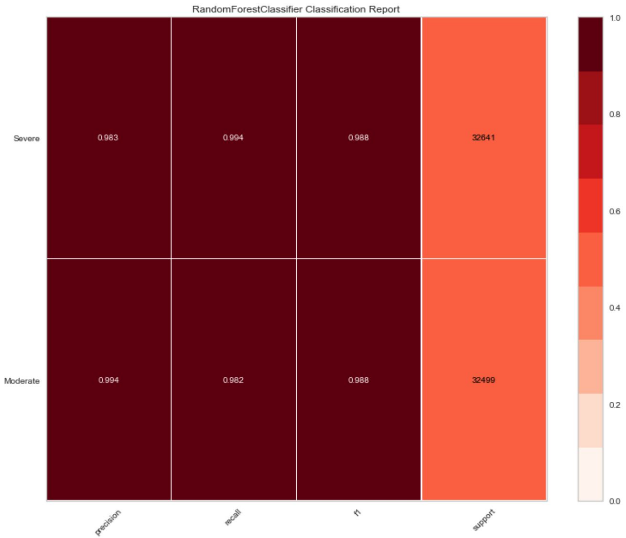


Massive class imbalance. Models slightly better than coin flip.

F1 Scores: .5 to .6

Feature Selection, **Modeling** and Tuning

- **SMOTE to the rescue.**
- Oversample our minority classes - Equal Bins



High Precision, Recall, F1
Suspicious?

- *What could go wrong?*
  - **Built the world’s best estimator to predict synthetic fires.**

Feature Selection, **Modeling** and Tuning

- **Fixes**
  - Don’t mix synthetic data and test data data.
  - Much more undersampling and a touch of oversampling.

- **Model training data is less imbalanced and better for training**

```
Counter({'Moderate': 1311429, 'Severe': 17493})
```

```
Counter({'Moderate': 150000, 'Severe': 35000})
```

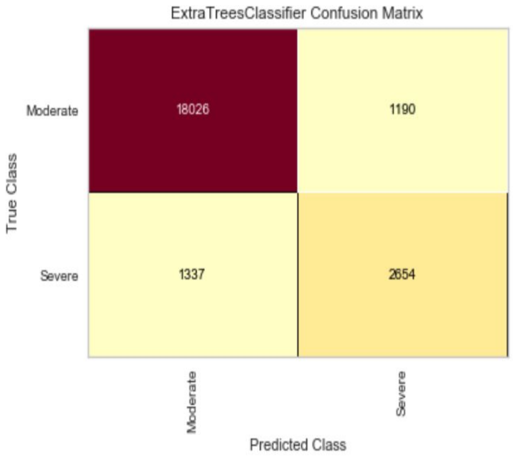
Feature Selection, **Modeling** and Tuning

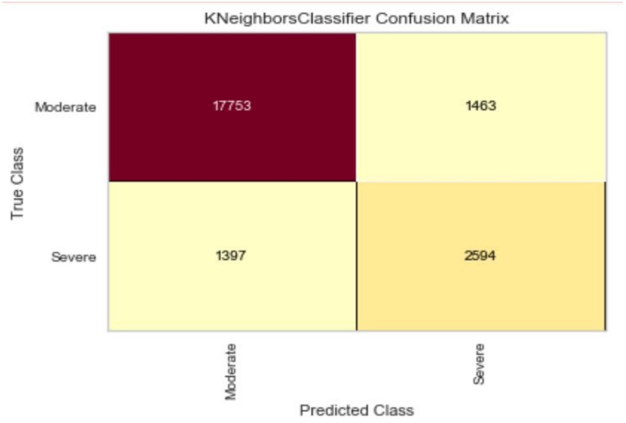
Initial Test Data

Models Evaluated	Findings	Further Evaluated
MultinomialNB()	Lower F1	No
LinearSVC()	Lower F1 -	No
KNeighborsClassifier()	Higher F1 / Slow	Yes
RandomForestClassifier()	Higher F1 -	Yes
ExtraTreesClassifier()	Higher F1 -	Yes
BaggingClassifier()	Slow, Complex	No
AdaBoost()	Lower F1, Poor Recall	No

Feature Selection, **Modeling** and Tuning

Moderate Fire Prediction is easy but what can predict severe?





- **Hyperparameters tuning a choice between precision and recall**
- GridSearch Employed but close the defaults worked best

Feature Selection, Modeling and **Tuning**

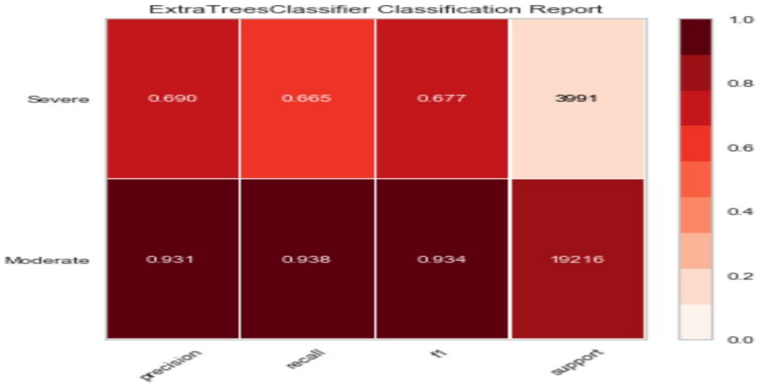
- **Models**
  - *Selected ExtraTreesClassifiers?*
  - StandardScaling() , OHE(),

- **Settled on High Recall to predict Severe fires though Low Precision.**

Feature Selection, Modeling and **Tuning**

Concerns: Overfitting, Class balance data, Unknowns **Positives:** Predictive of most dangerous fires (better than a chance)

Consider: Other or Multiple Classifiers - KNeighborsClassifier and Randomforest were close.



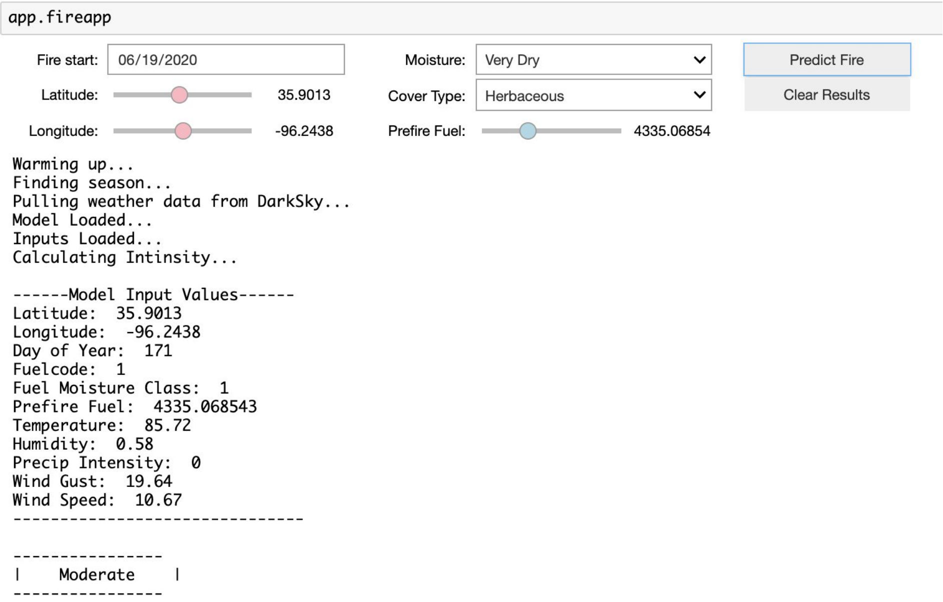

```
Pipeline : Original data F1 Micro Score : 0.8911104408152711
Pipeline : Original data F2 Macro Score : 0.8059854605438945
              precision    recall  f1-score   support

    Moderate       0.93      0.94      0.93     19216
      Severe       0.69      0.66      0.68      3991

    accuracy                           0.89     23207
   macro avg       0.81      0.80      0.81     23207
weighted avg       0.89      0.89      0.89     23207
```

Conclusion

- Lessons Learned
- If we had time…
- Class balance matters in Classification!



Thank you