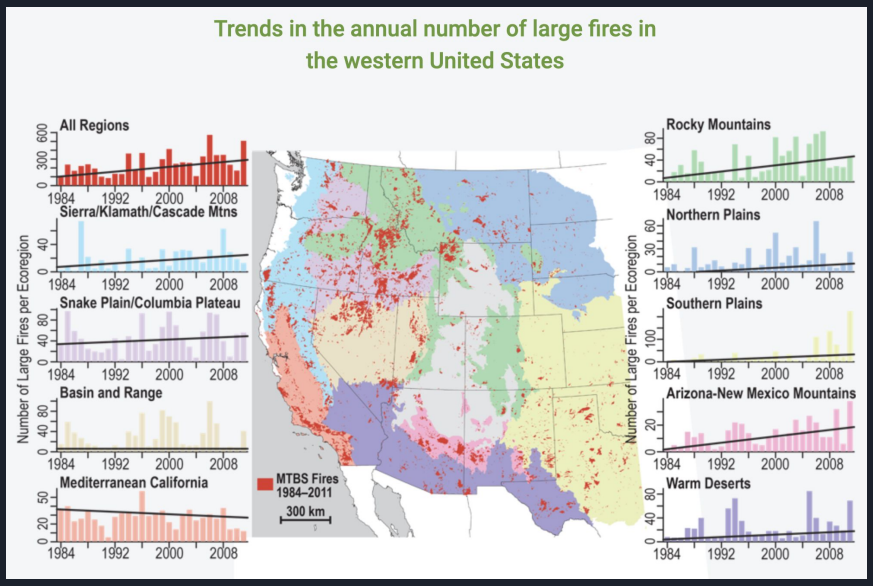
Predict Wildfire Spread

Mike Freel
Jiselle Jones
John-Paul Kivlin

JP

Reason for topic



[Center for Climate and Energy Solutions]

This graphic from the Fourth National Climate Assessment shows the growth in large wildfires throughout the West. The black lines are fitted trend lines. Statistically significant at a 10% level for all regions except the Snake Plain/Columbian Plateau, Basin and Range, and Mediterranean California regions.

SOURCE
[Dennison et al]

Jiselle
According to the Center for Climate and Energy Solutions, wildfires in the Oregon region (and many other places across the United States) seem to be increasing over time in both frequency and intensity. These wildfires impact all life in surrounding areas. Not only are the increasing trends of wildfires caused by climate change, they also impact and are accelerating Earth's changing climate and species extinction. If we can predict the location, size, and spread of a wildfire, we may be able to decrease the amount of life lost as well as slow the increasing global temperatures.

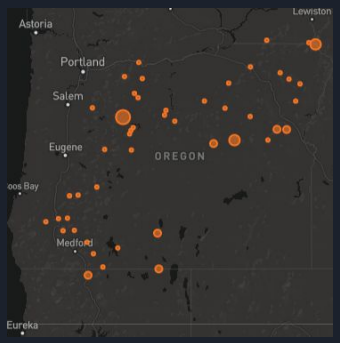
Purpose

Predict Wildfire Spread

Fuel Type

*Photo by Sara Farshchi on Unsplash

Location



Weather

*Photo by Jonathan Bowers on Unsplash

Jiselle

Technologies

Database

- postgreSQL
- pgAdmin
- Heroku pipeline

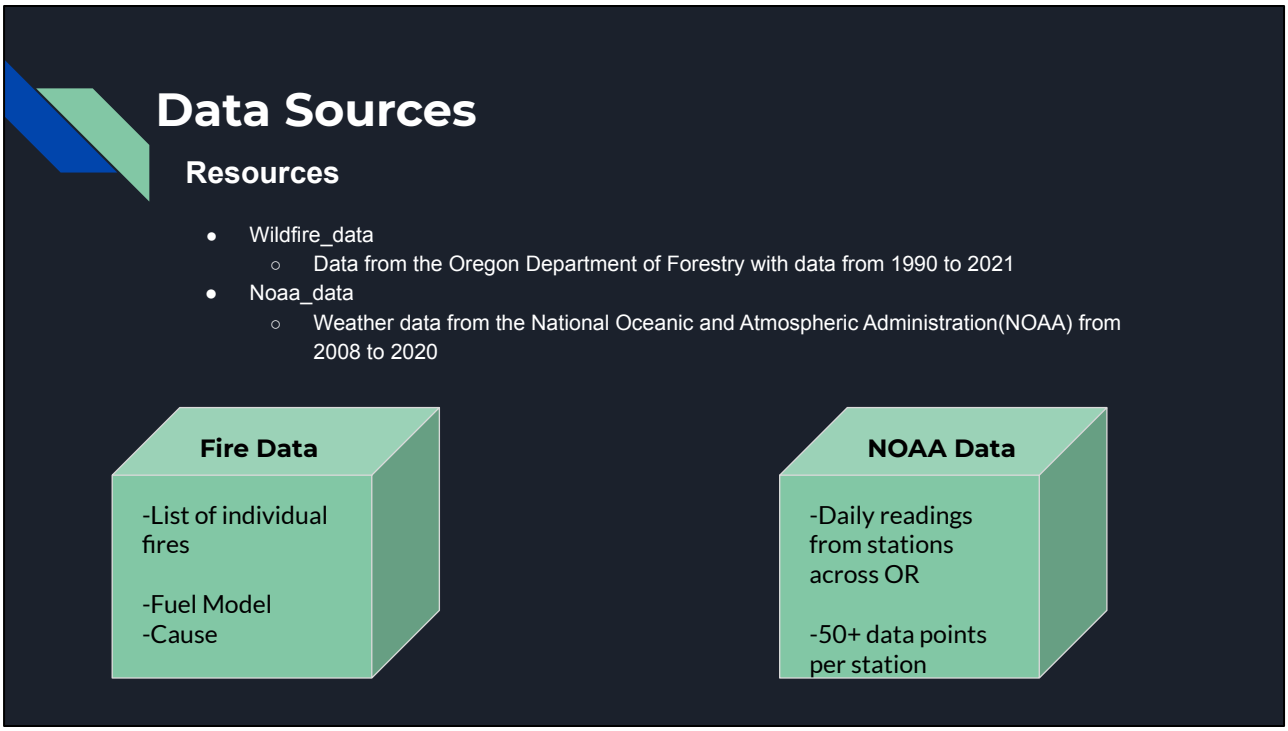
Exploratory Data Analysis

- postgreSQL
- SQLALchemy
- Pandas
- Python 3.7
- VSCode 1.62.0
- PGAdmin 5.2
- Scikit learn

Dashboard Technology

- Flask
- HTML
- Javascript
- Leaflet
- Highcharts
- JSON
- Heroku deployment

JP
Quick overview of the technologies we used

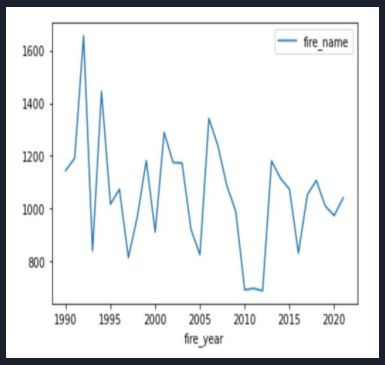


Fire we got the weather data
After we started compiling that, we hunted for the weather data, which we found from NOAA
Neither data source was easy to get data from

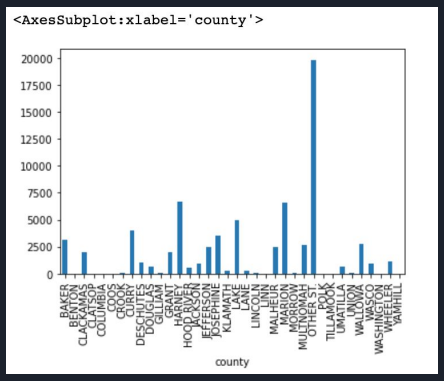
- The Fuel model categorizes grasses, shrubs, trees, dead leaves, and fallen pine needles. As these burnable materials pile up, so do the chances of catastrophic wild-land fire

Exploratory Data Analysis
All Fire Data 1990 - 2021

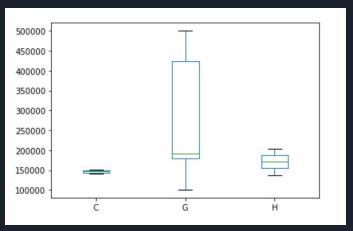
of Fires Over Time



Of Fires per County



Fuel Type - Large Fire



Fuel Models
Fuel Model C: Open pine, grass under
Fuel Model G: Conifer, Old growth
Fuel Model H: Conifer, Second growth

Jiselle
In preliminary analysis we wanted to get a picture of the data. What does the wildfire data look like in Oregon?

** click arrow
We started by looking at the number of fires over time which seemed to be decreasing, but, as we saw earlier, the severity of fires is actually increasing, so what could be causing the increase in severity?

**click arrow
We decided to then look at the places where most fires occur and found that there have been 5,000 or more fires in Harney County, Lake County, and Marion County between 1990 and 2021. We wondered what the ecosystem is like in those areas. What type of vegetation is there. Could it be that the type of vegetation or fuel is correlated with fire severity?

**click arrow
Then we began considering the fuel type in each area. Looking at a box plot of all fires was really difficult to see, so we decided to focus on the really severe fires, those that burned 100,00 acres or more and found that there are 3 types of vegetation present in those fires...see table

Exploratory Data Analysis - weather

- Dropped all columns from the weather data except for the 5 core categories as tracked by noaa

The five core values are:

PRCP = Precipitation (mm or inches as per user preference, inches to hundredths on Daily Form pdf file)
SNOW = Snowfall (mm or inches as per user preference, inches to tenths on Daily Form pdf file)
SNWD = Snow depth (mm or inches as per user preference, inches on Daily Form pdf file)
TMAX = Maximum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file
TMIN = Minimum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file

JP

Machine Learning Model - Data Processing

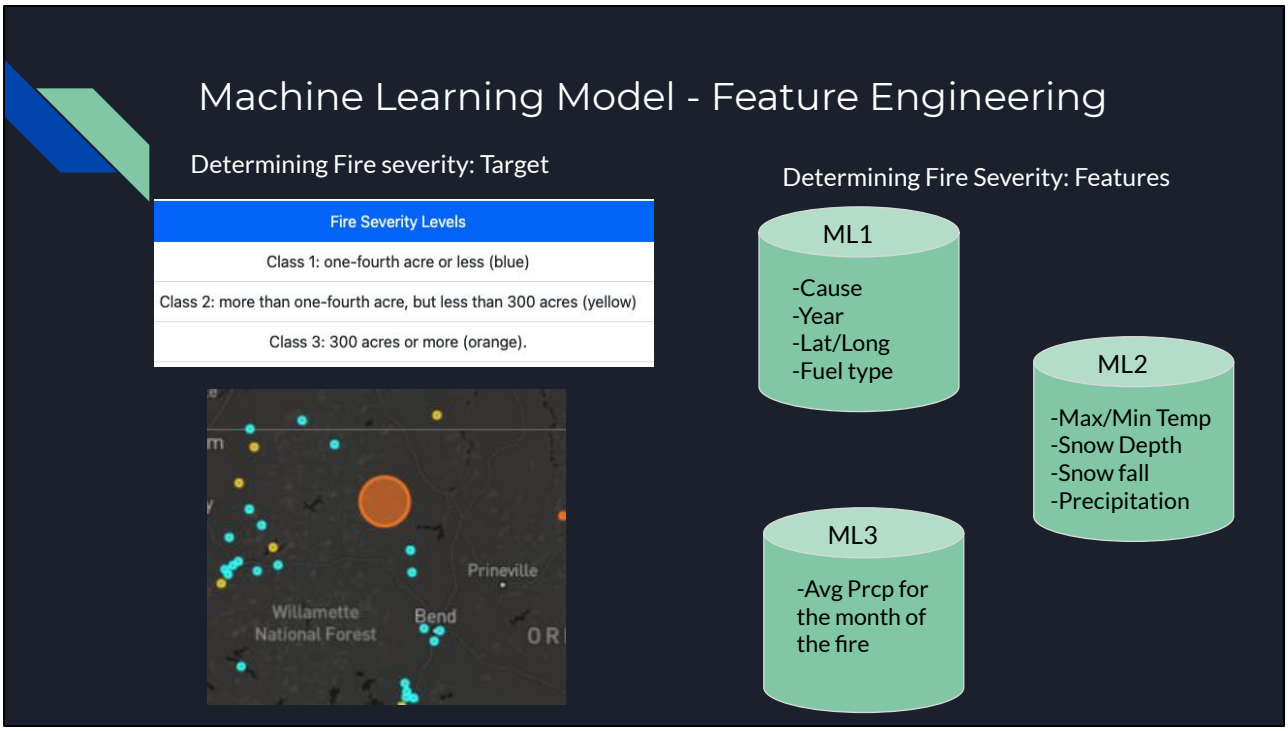
Data Cleaning Process:

- Change Fire Year data to 2008 -2020 to match weather data
- Removal of null values - removed ~10% after merge
- Data Type: Converted Lat/Long from DMS to Decimal
- Label Encoder: Change fuel type from letter to numbers
- Severity levels binned 2 different ways
- General Cause:

```
# Change General Cause Human or Nature

fire_df['general_cause'] = fire_df['general_cause'].replace(['Recreationist','Equipment Use','Debris Burning', 'Smokin
fire_df['general_cause'] = fire_df['general_cause'].replace(['Lightning'],'2')
fire_df['general_cause'] = fire_df['general_cause'].replace(['Under Invest', 'Miscellaneous'],'3')
```

JP
General Cause (human) - 'Recreationist','Equipment Use','Debris Burning', 'Smoking', 'Arson', 'Railroad', 'Juveniles



JP

Determining Fire Spread: Target

For our target, we determined there was value in following a recognized fire classification system. The levels were changed to match how the National Wildfire Coordinating Group classifies fires. After running the model with the changed class sizes and adjusting the years 2008 -2020 (data loss), we saw a decrease in accuracy across the model . Therefore, the class sizes were adjusted to follow a 3 class system from USDA Forest Service.

Classification is based on the total amount of acres burned

Talk

TRAINING/TESTING

Random Forest

```
# Define the features set.
X = fire_binary_encoded
X = X.drop("fire_severity", axis=1)
X.head()
```

```
# Define the target set.
y = fire_binary_encoded["fire_severity"].ravel()
y[:5]
```

```
# Splitting into Train and Test sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=78)
```

SMOTEENN

```
x_cols = [i for i in fire_smoteen.columns if i not in ('fire_severity')]
X = fire_smoteen[x_cols]
y = fire_smoteen['fire_severity']
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1, stratify=y)
```

Jiselle

Scaling/Fitting Random Forest

```
# Creating a StandardScaler instance.
scaler = StandardScaler()
# Fitting the Standard Scaler with the training data.
X_scaler = scaler.fit(X_train)

# Scaling the data.
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```

```
# Create a random forest classifier.
rf_model = RandomForestClassifier(n_estimators=100, random_state=78)
```

```
# Fitting the model
rf_model = rf_model.fit(X_train_scaled, y_train)
```

Make Predictions
Used this for Map & Graph in Dashboard

```
predictions = rf_model.predict(X_test_scaled)
```

Jiselle
Estimators = 100 (trees)

Resampling/Fitting SMOTEEN

```
from imblearn.combine import SMOTEENN
smoteenn = SMOTEENN(random_state=1)
X_resampled, y_resampled = smoteenn.fit_resample(X_train, y_train)
Counter(y_resampled)
```

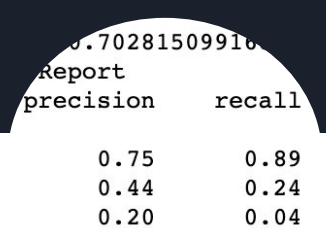
```
from sklearn.linear_model import LogisticRegression
smoteen_model = LogisticRegression(solver='lbfgs',max_iter=100)
smoteen_model.fit(X_resampled, y_resampled)
```

Make Predictions

```
from sklearn.metrics import confusion_matrix
y_pred = smoteen_model.predict(X_test)
confusion_matrix(y_test, y_pred)
```

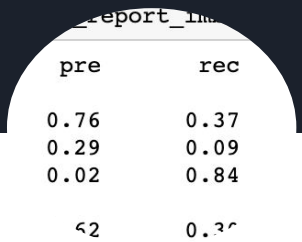
Jiselle
Max_iter of 100

Machine Learning Model 1 - Fire Data Only



Random Forest Classifier

Acc = 70%



SMOTEENN

Acc = 43%

The module talks about the benefits of the Random Forest Classifier.
The first thing it talks about are how RFC is robust against overfitting because weak learners are trained on different pieces of the data. But one of the complications we ran into was that we had thousands of fires at severity level 1 and 2 and far fewer at severity level 3.
Let me show you on our dashboard--- ***show dashboard
(47 level 3, over 1000 levels 1 & 2)

You can see that we have xxx which, as JP mentioned before is only about 4% of the data. So our model could only train on a very small set of the data and this is why we decided to try SMOTEENN, but when we ran the model, the overall accuracy was only 43%, which was far less than the RFC.

RFC can also handle thousands of input variables without variable deletion. If we were able to include more variables like drought conditions, sea water temperature, etc, then this would be beneficial going forward.

Also, RFC's are robust to outliers which we have in our data set since many of the larger fires were considered outliers when considering all data together.

In our initial run of models, we really honed in on the accuracy of the model, but in retrospect, we might also have considered the precision and recall.

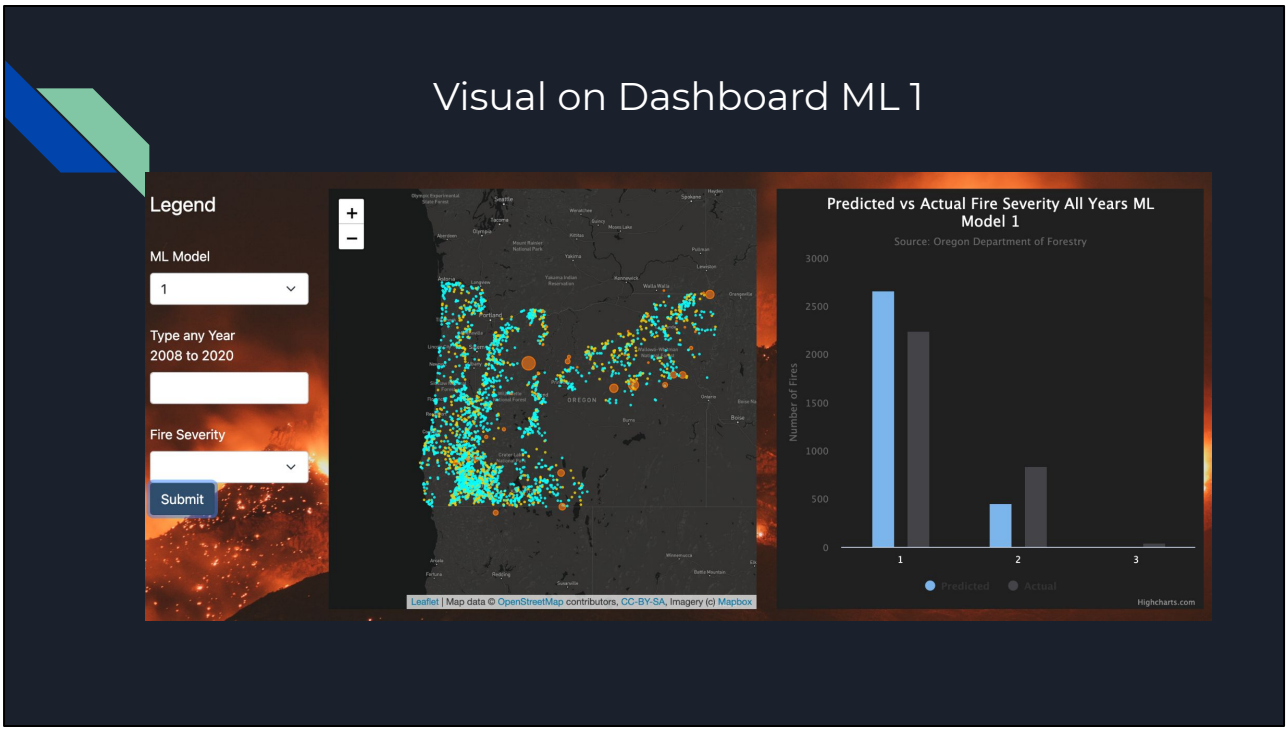
While the accuracy of the RFC was 70%, this could have been due to the large

number of level 1 fires we had, so we think it was overfitting the data.

When you look at recall on level 3, the RFC is only at 4% which is really poor. But the SMOTEEN is 84%. This is something we could investigate further.

Just to show what I mean visually Show dashboard graph

Informational if it comes up
Recall = (True pos/predicted results) (ROW)
Precision - How many did we get right? (true pos/actual results) (COLUMN) Precision is going to be higher in the severity level 3 because we have such a small number of level 3 fires to test (47 for RFC & 32 for SMOTEENN)

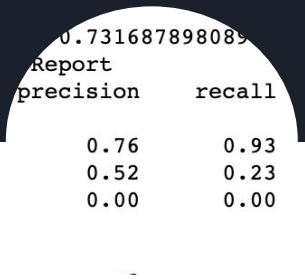


Visual on Dashboard

Jiselle
Limitations: Potential for overfitting (class 1), high chance the trees had their own circumstances, like class imbalance, sample duplication and wrong node splitting.
Iteration
% of fires that were severity 1 ()

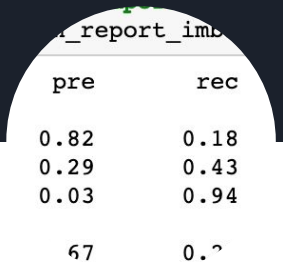
ML1 - 70% accuracy score and only model to predict class 3 fire severity - weighted = everything, macro =
Ml2- 73% accuracy, weather data added could not detect class 3 fires
Ml2 - 71% accuracy, decrease in accuracy and unable to predict class 3 fires. Attempted SMOTEENN to account for overfitting.

Machine Learning Model 2
Fire Data + Weather on Fire Start Date



Random Forest Classifier

Acc = 73%



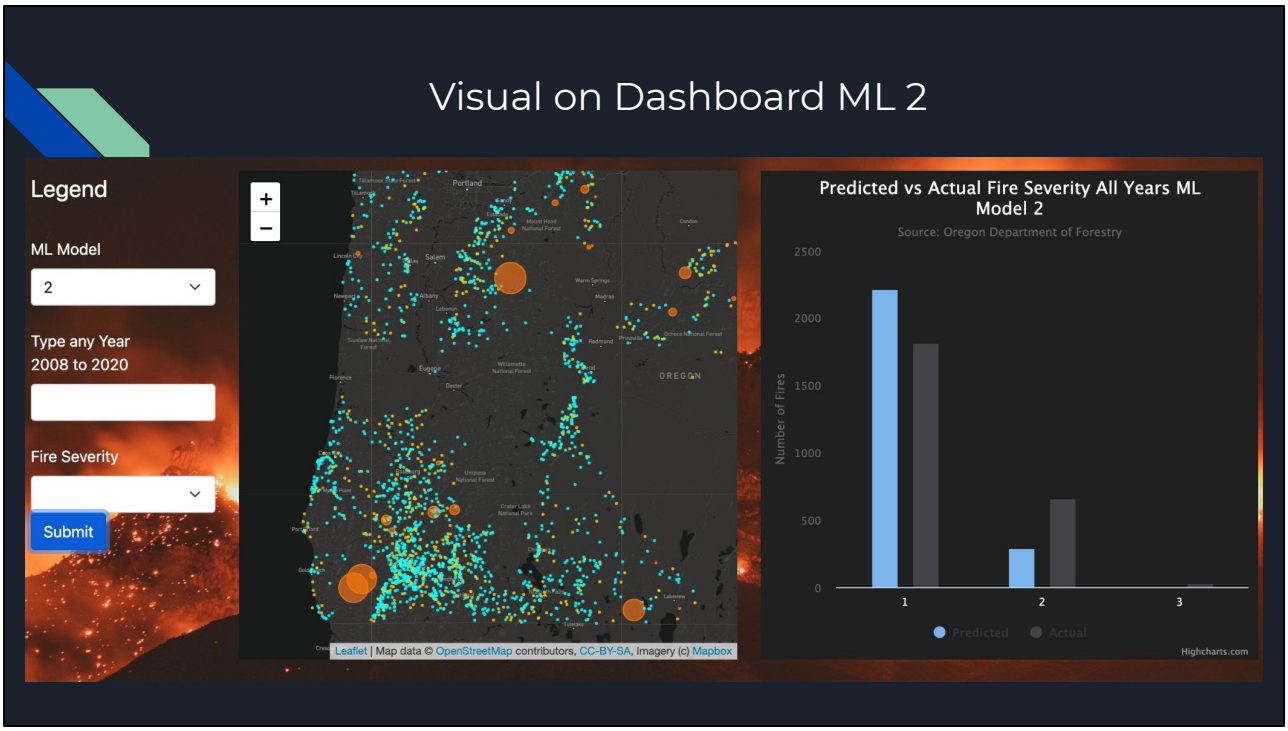
SMOTEENN

Acc = 52%

Jiselle

Limitations: Potential for overfitting (class 1), high chance the trees had their own circumstances, like class imbalance, sample duplication and wrong node splitting.
Iteration
% of fires that were severity 1 ()

ML1 - 70% accuracy score and only model to predict class 3 fire severity - weighted = everything, macro =
Ml2- 73% accuracy, weather data added could not detect class 3 fires
Ml2 - 71% accuracy, decrease in accuracy and unable to predict class 3 fires. Attempted SMOTEENN to account for overfitting.



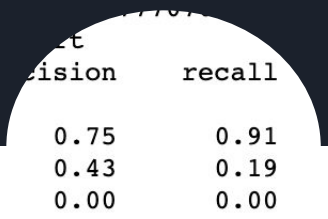
Visual on Dashboard

Jiselle
Limitations: Potential for overfitting (class 1), high chance the trees had their own circumstances, like class imbalance, sample duplication and wrong node splitting.
Iteration
% of fires that were severity 1 ()

ML1 - 70% accuracy score and only model to predict class 3 fire severity - weighted = everything, macro =
Ml2- 73% accuracy, weather data added could not detect class 3 fires
Ml2 - 71% accuracy, decrease in accuracy and unable to predict class 3 fires. Attempted SMOTEENN to account for overfitting.

Machine Learning Model 3

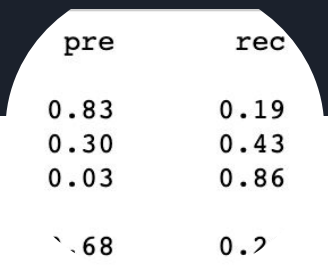
Fire + Weather Data on Fire Start Date + Avg Precipitation Month



	cision	recall
	0.75	0.91
	0.43	0.19
	0.00	0.00

Random Forest Classifier

Acc = 71%

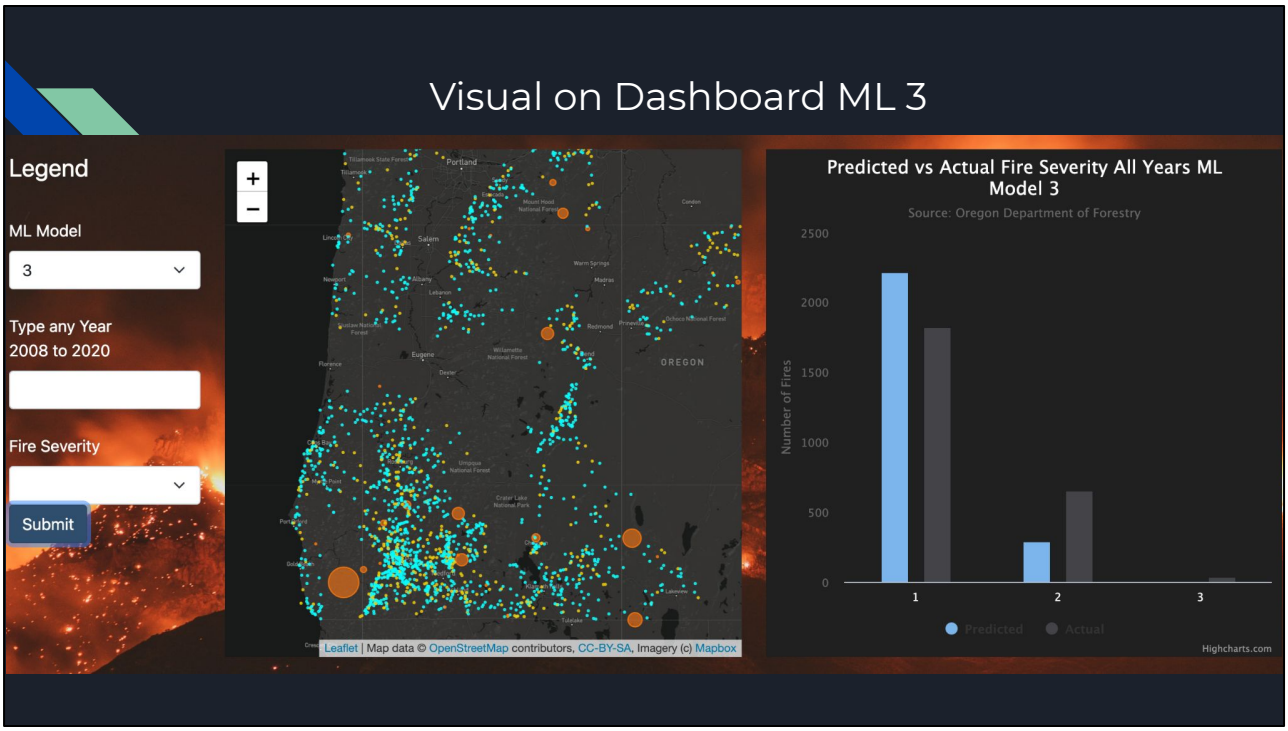


	pre	rec
	0.83	0.19
	0.30	0.43
	0.03	0.86
	.68	0.2

SMOTEENN

Acc = 49%

Jiselle
Limitations: Potential for overfitting (class 1), high chance the trees had their own circumstances, like class imbalance, sample duplication and wrong node splitting.
Iteration
% of fires that were severity 1 ()



Visual on Dashboard

Feature Importance

```
# Calculate feature importance in the Random Forest model.
importances = rf_model.feature_importances_
importances
```

ML 1

	label character varying (252)	importance numeric
1	Longitude	38.2
2	Latitude	37.2
3	fire_year	14.5
4	fuel_model_X	1.5
5	fuel_model_A	0.8
6	fuel_model_C	0.8
7	fuel_model_L	0.7
8	general_cause_1	0.7
9	fuel_model_H	0.7
10	general_cause_2	0.6
11	fuel_model_F	0.6
12	fuel_model_J	0.6
13	general_cause_3	0.5
14	fuel_model_G	0.5

ML 2

	label character varying	importance numeric
1	Longitude	18.7
2	Latitude	18.0
3	tmin_avg	15.2
4	tmax_avg	14.6
5	fire_year	8.6
6	prcp_avg	8.1
7	snwd_avg	2.3
8	fuel_model_X	1.5
9	fuel_model_A	1.4
10	fuel_model_C	1.3
11	fuel_model_H	1.3
12	fuel_model_L	1.2
13	general_cause_1	1.2
14	fuel_model_F	0.9

ML 3

	label character varying	importance numeric
1	Longitude	18.0
2	Latitude	17.1
3	tmin_avg	14.4
4	tmax_avg	13.9
5	avg_prcp	12.3
6	fire_year	7.7
7	snwd_avg	2.2
8	fuel_model_X	1.4
9	fuel_model_A	1.4
10	fuel_model_H	1.3
11	fuel_model_C	1.2
12	general_cause_1	1.2
13	fuel_model_L	1.2
14	general_cause_2	0.9

Jiselle
Another thing the RFC can be used for is to rank the importance of input variables in a natural way. We'll look at our Feature Importances later.

Legend For Fuel Models
#A Annual grasses (cheat)
#B Dense Chaparral
#C Open pine, grass under
#F Dense Brush (lighter than B)
#G Conifer, Old growth
#H Conifer, Second growth
#I Slash, heavy
#J Slash, medium
#K Slash, thinning, P.C., Scattrd
#L Grass Perennial
#R Hardwood, summer
#T Sagebrush, medium dense
#U Closed canopy pine
#X Non wildland fuel

Legend For General Cause
1 = Human
2 = Nature
3 = Uncategorized

Future Analysis Recommendation

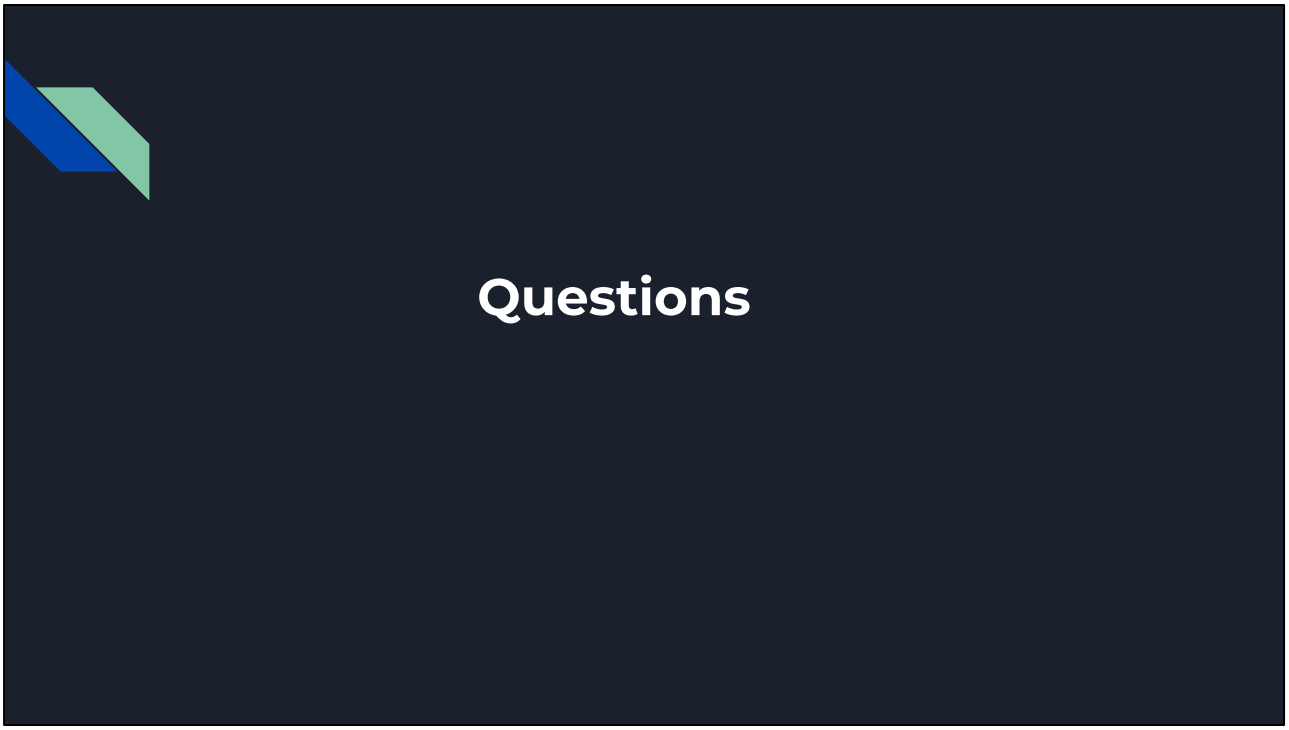
Explore weather data further

- Bin/classify better weather data
  - Averages of snowpack, etc. where applicable
- Further research on Sea water temperature
- Aggregate weather data by last 60/90 + days
- Weather data with more features
- Add more years of weather data
- Drought conditions

Machine Learning

- Investigate recall on SMOTEENN further
  - Scale data on SMOTEEN to see if we get better results
- Further analysis on data we did get in each model
- Run a binary model on severity level 3 only
  - Can we predict a level 3 fire? yes/no

JP/Jiselle



Jiselle