US Wildfire Scale Prediction

Team2: Tzu-Yao Lin, Shan Xiang,

Bodong Xu, Ningyi Xue

Introduction

Dataset Info:

- Wildfires are among the most common form of natural disaster. Last year, more than four million acres have burned by wildfires.
- 1.88 million wildfire records
- 39 features

What kind of question we want to explore:

- How is US wildfires geographically distributed from 1992 to 2015?
- What starting conditions are associated with wildfire size?

Purpose of our study

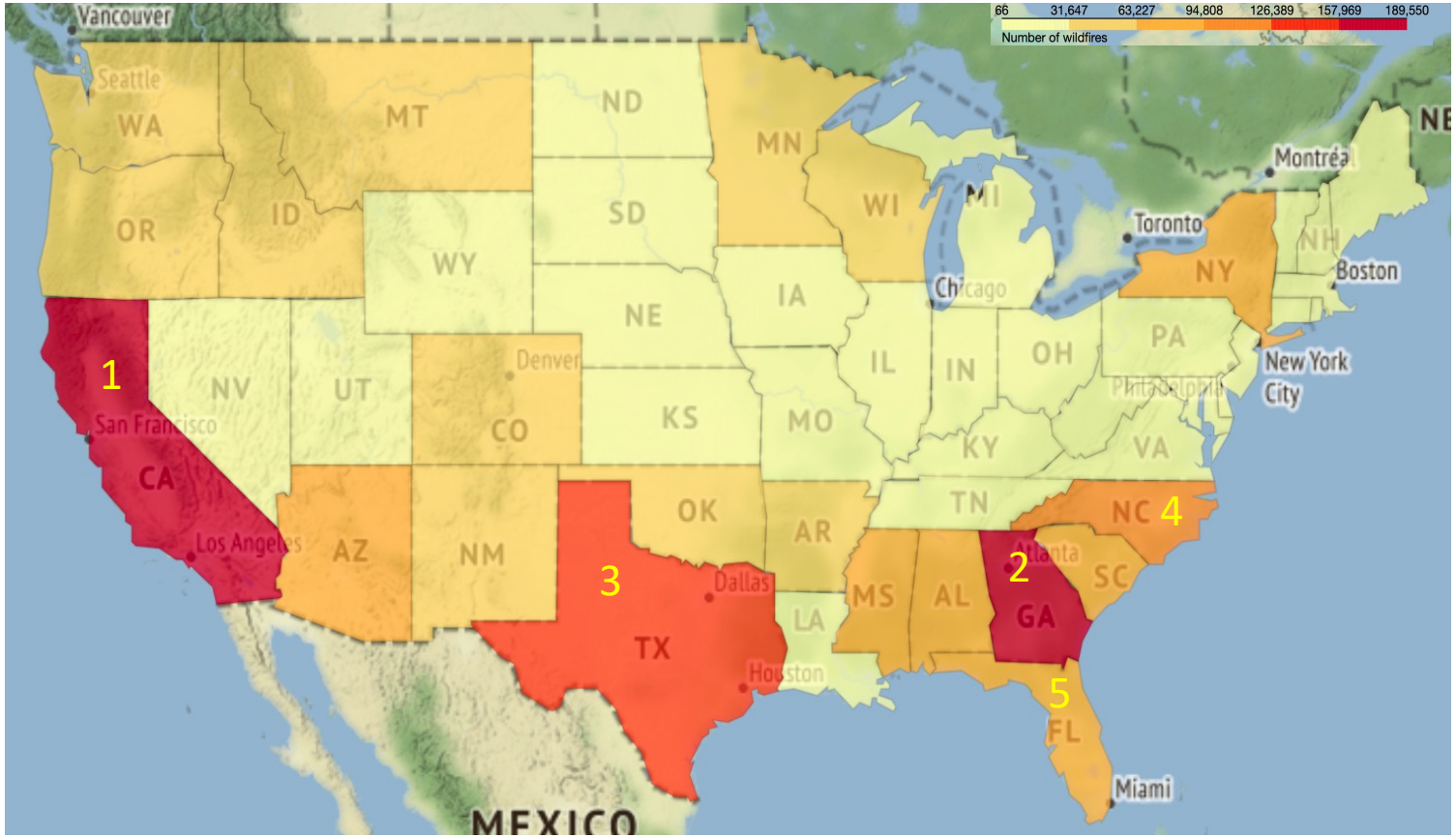
- Predict the size of wildfire from basic starting conditions
- Better prediction for more effective actions to be taken in advance

Exploratory Data Analysis

Number of Wildfires Over States between 1992 - 2015

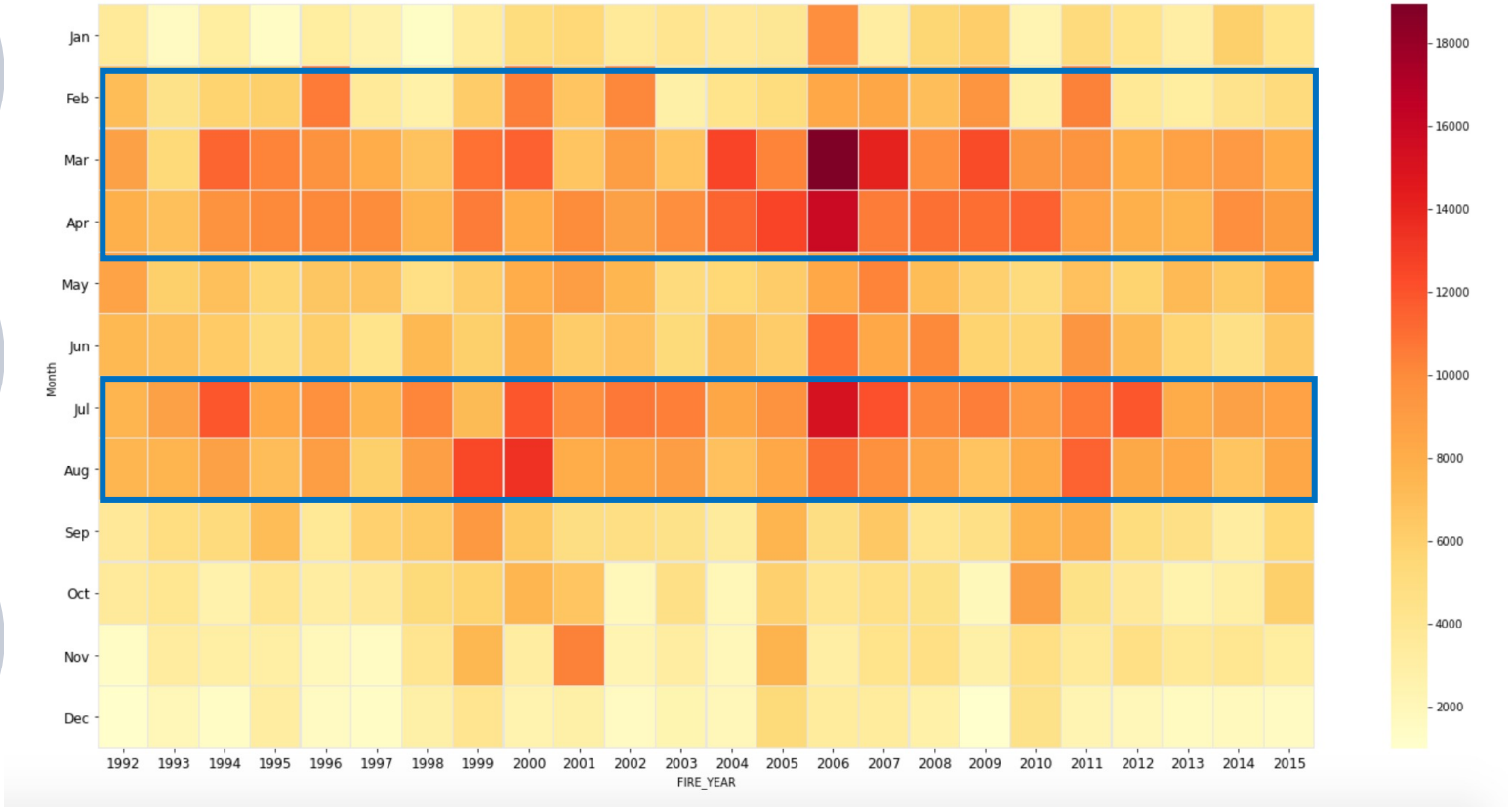
TOP 5 State:

1. CA
2. GA
3. TX
4. NC
5. FL



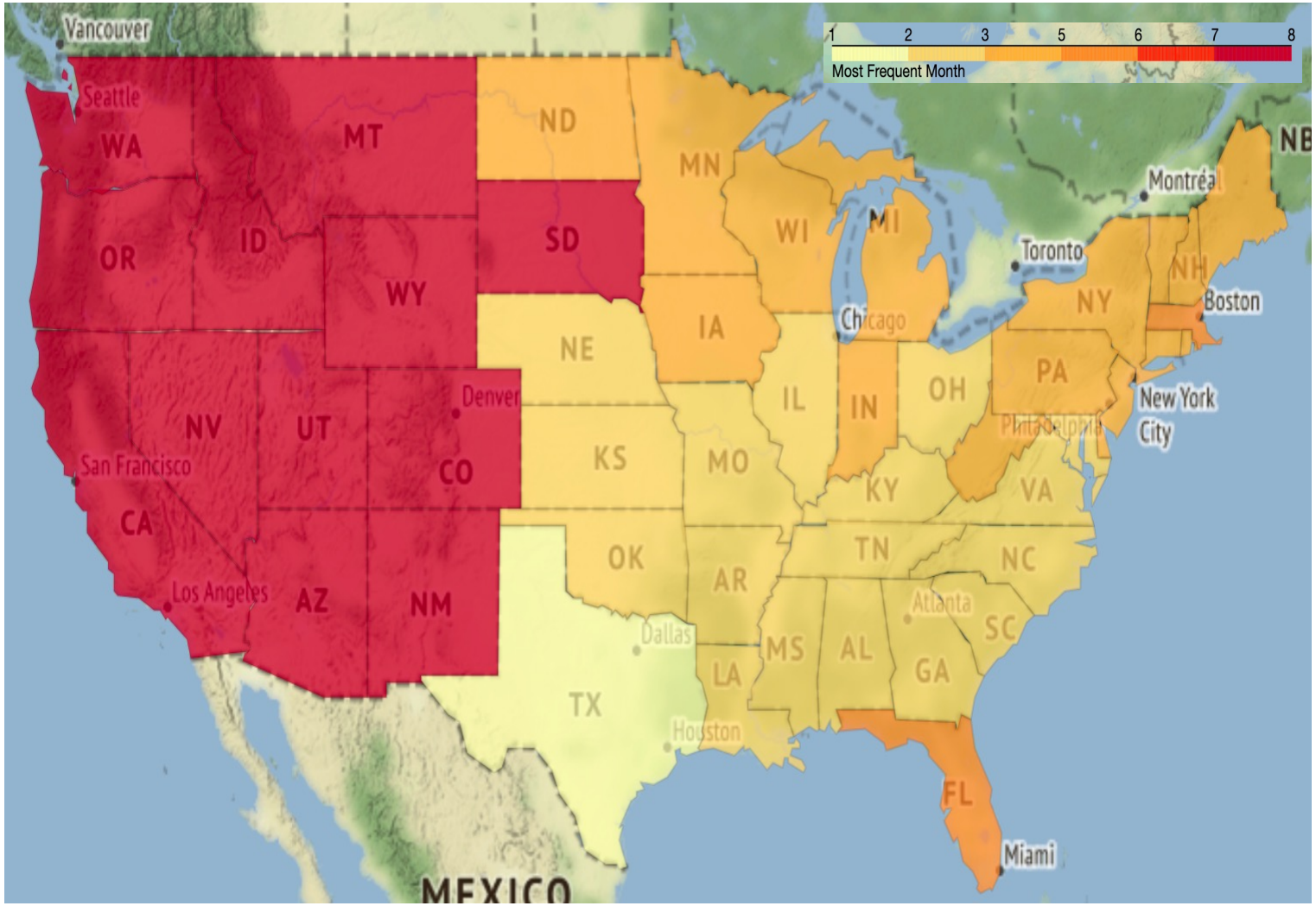
US Wildfire Year and Month Heatmap

- March 2006 is the most severe month crossing all year and moth



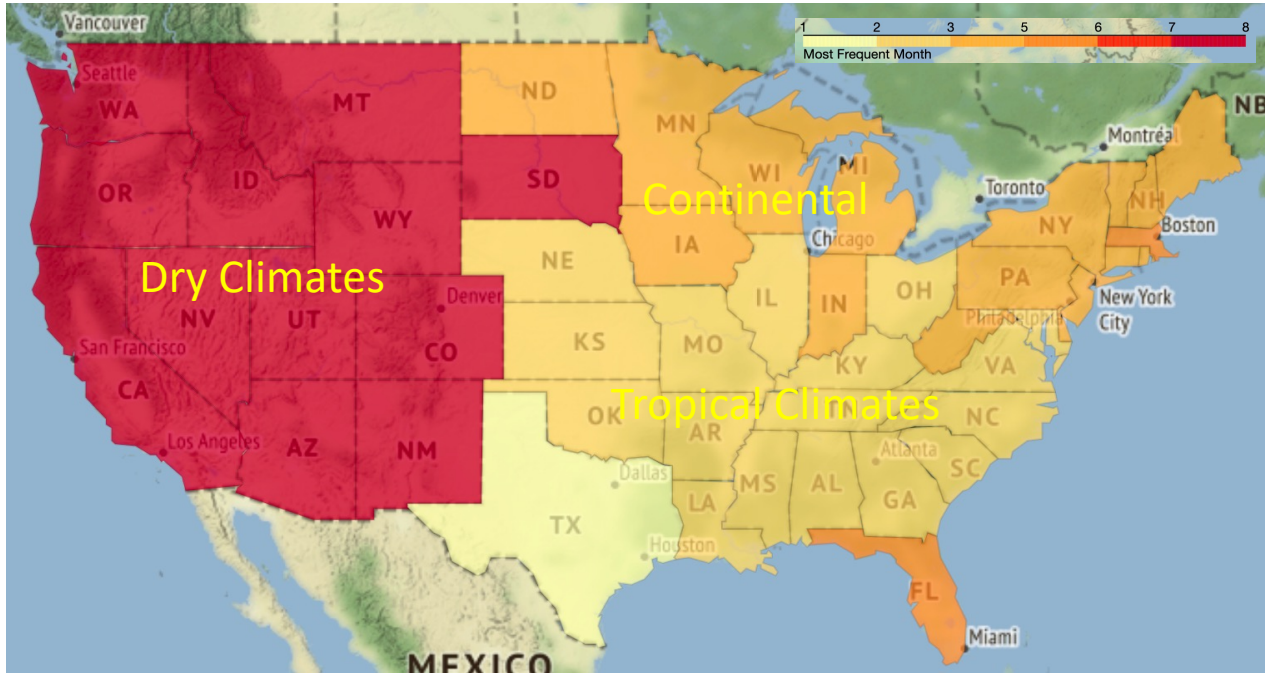
US Wildfire Most Frequent Month

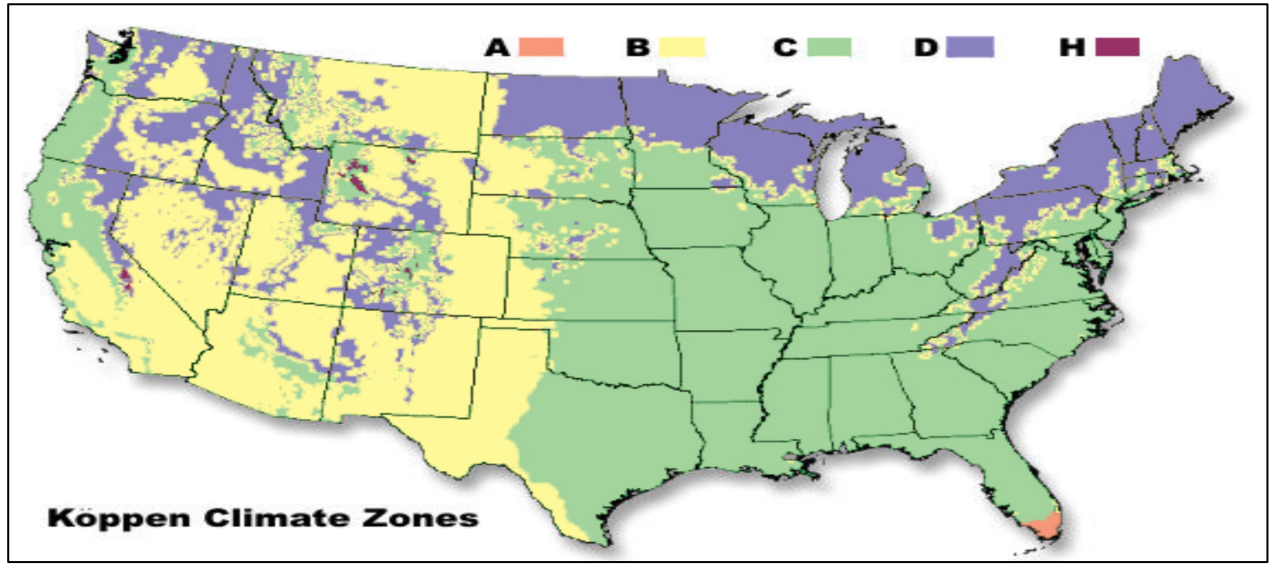
- West : August
- North : March
- Center : February



US Wildfire Most Frequent Month

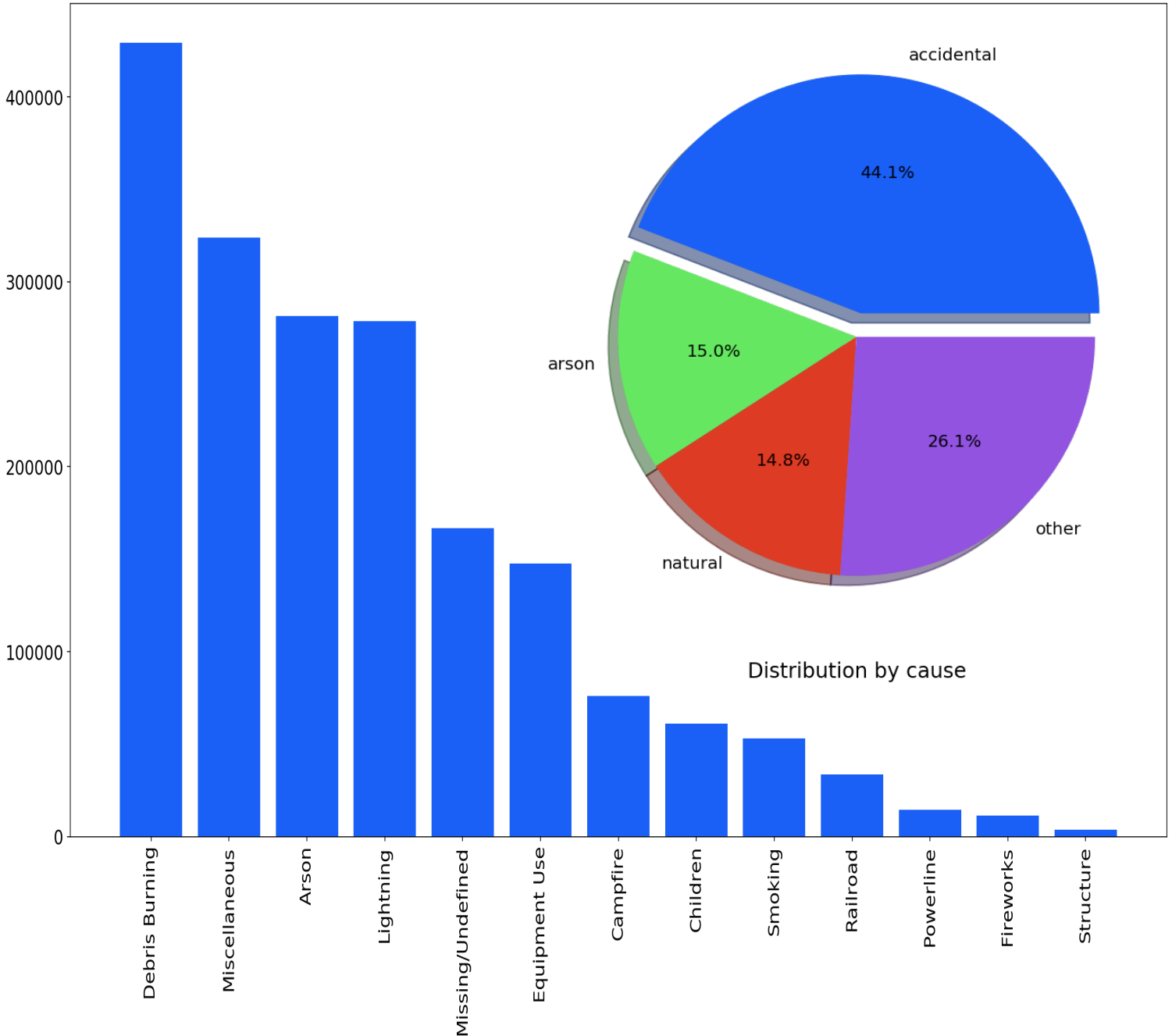
- West : Dry
- North : Continental
- Center : Temperate





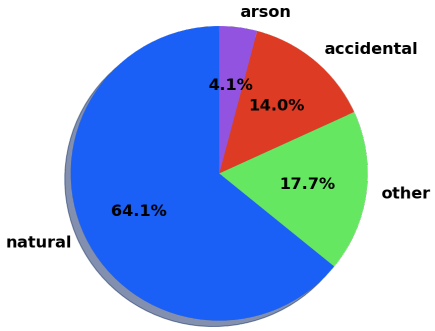
US Wildfire Cause Analysis

- Natural : Lightning
- Arson : Arson
- Accidental : Debris Burining, Equipment use, Campfire,etc.
- Other : Miscellaneous, Missing

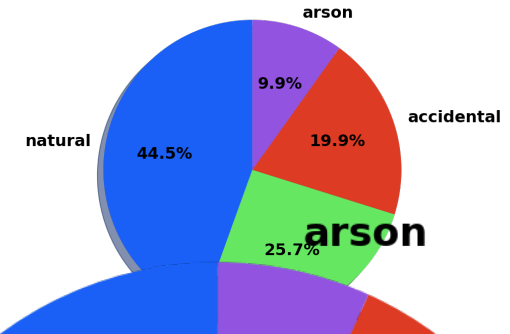


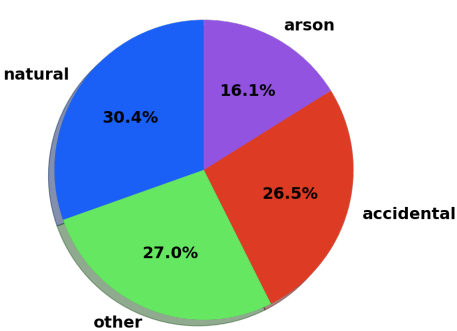
US Wildfire Cause Analysis by Fire Size

- 62% burned area was caused by lightning
- Above 300 acres, Lightning is a major factor
- Below 300 acres, accidental is a major factor

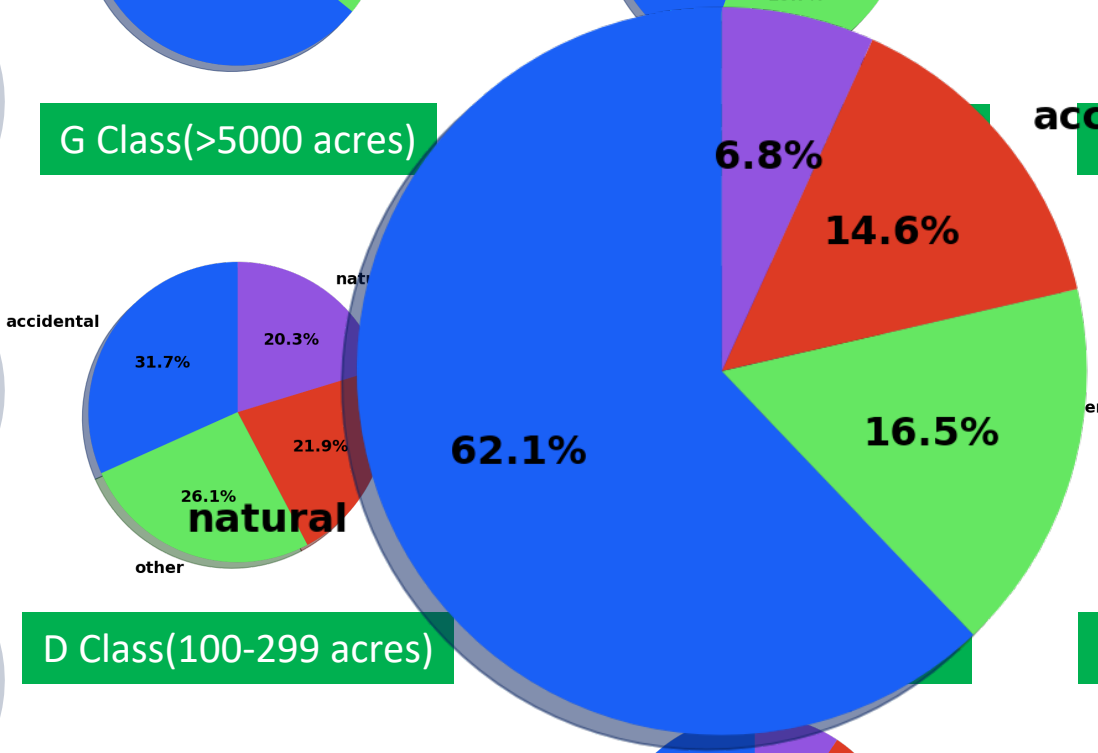


G Class(>5000 acres)

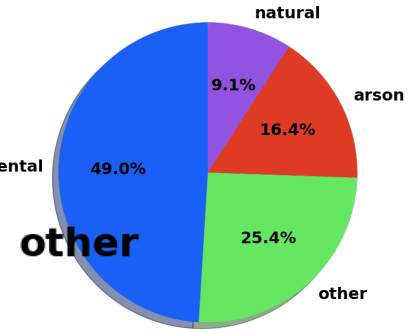




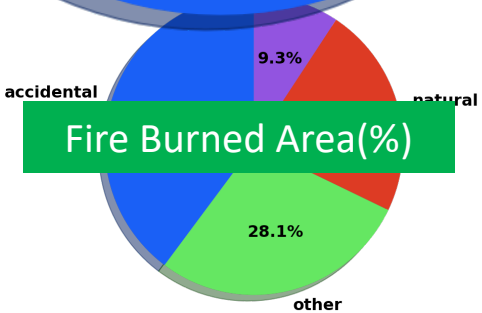
E Class(300-999 acres)



D Class(100-299 acres)



B Class(0.26-9.9 acres)

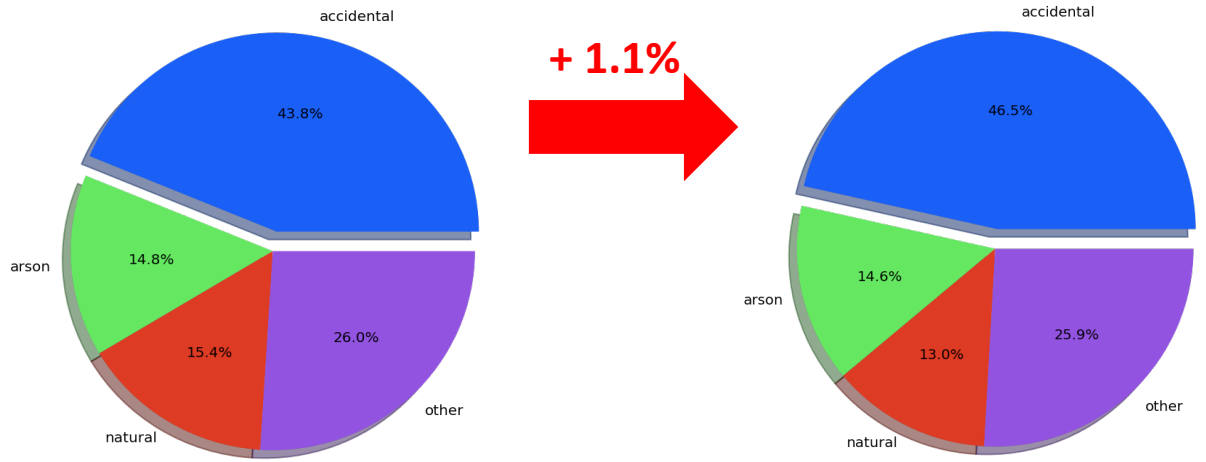


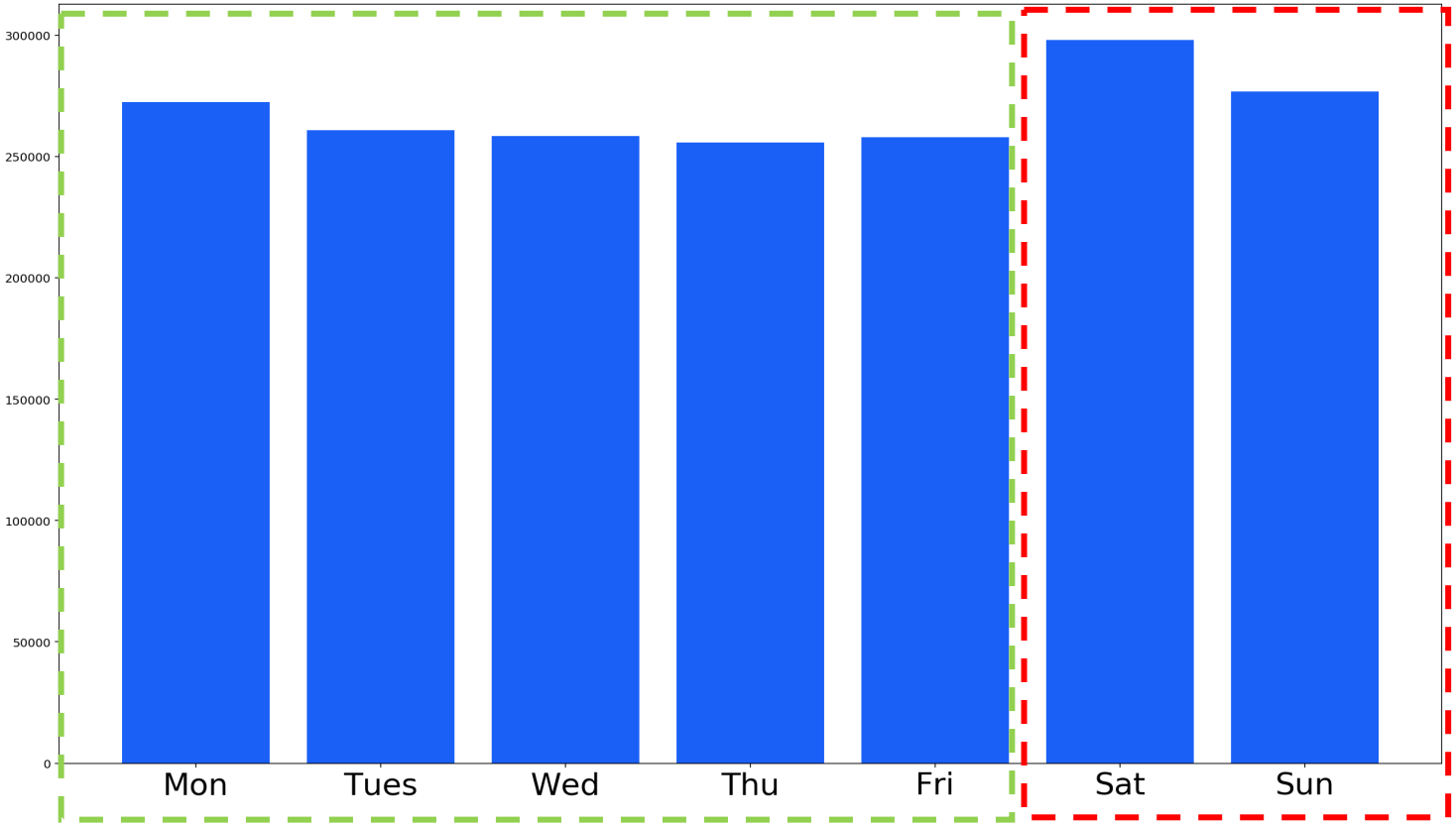
Fire Burned Area(%)

A Class(<0.25 acres)

US Wildfire Analysis by day of week

- More fires on weekend
  - Accidental + 1.1%
  - Careless camp fires, smoking, fireworks





Modeling and Evaluation

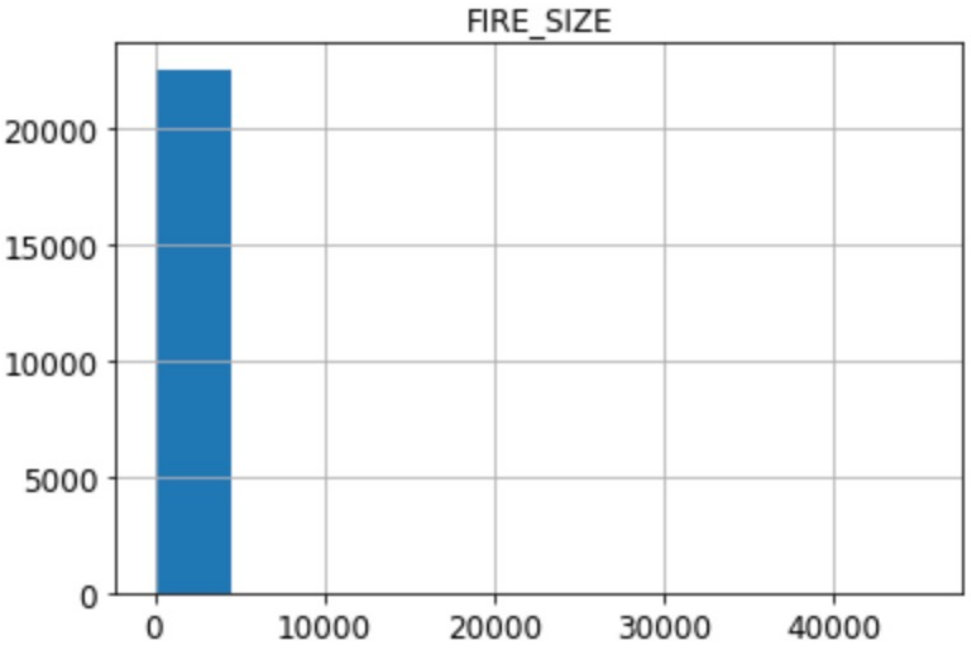
Regression Analysis

Improved Classification

Hyperparameter Tuning

Predicting Objective

Regression- Area within the final perimeter of the fire



Mean: 14.6 Median: 0.5 Max：45294

Classification- Code for fire size based on the number of areas within the final fire

Regression Analysis

We initially try to predict the area within the final perimeter of the fire

Features we select:

- STAT_CAUSE_DESCR: 12 listed causes of the fire
- FIRE_YEAR: Year in which the fire was discovered
- DISCOVERY_DOY: Day of year on which the fire was discovered
- DISCOVERY_TIME: Time of day that the fire was discovered
- Dayofweek: Day of week on which the fire was discovered
- LATITUDE
- LONGITUDE

Correlation:

Regression Model Comparison

Model We Use

Linear Regression
Decision Tree
Random Forest
Support Vector Machine
KNN

Evaluation Matrix

Mean Squared Error
Cross validation MSE
Cross validation Standard Deviation

Result

Overfitting

	Linear Regression	Decision Tree	Random Forest	SVR	KNN
RMSE Score	482.24	0.07	187.00	483.71	438.53
Mean score of cross validation	352.96	504.95	423.64	349.47	463.80
Std of cross validation	329.98	420.05	291.81	334.42	268.39

Improved Classification

Model we use:

Logistic Regression
Decision Tree
Random Forest
Support Vector Machine
KNN

How we improve the prediction model

Use fire size class instead of fire size

Group 12 causes of fires into four categories

- Accidental
- Arson
- Natural
- Others

- Compare models using 12 causes and 4 categories
- Select best-performed model to do parameter tuning and prediction using testing data

Classification Comparison

- Evaluation matrix is accuracy
- Different methods to categorize causes of method don’t significantly improve the performance
- **Random forests** perform the best among all other models

12 original causes of fire

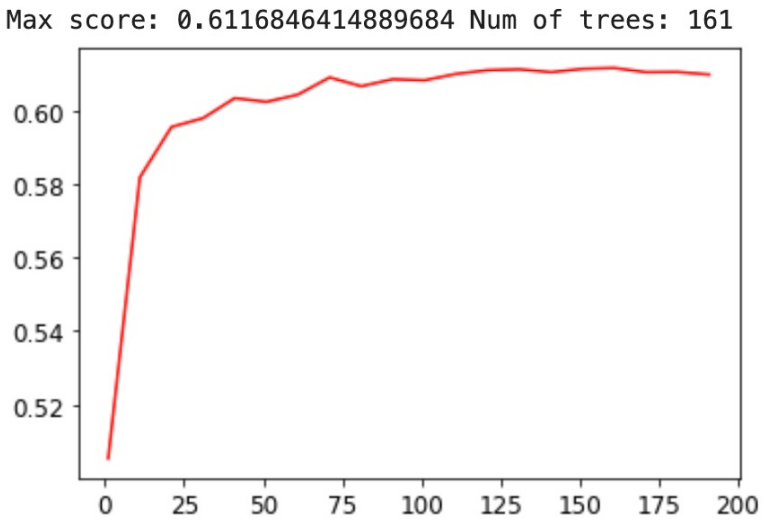
	Logistic Regression	Decision tree	SVM	KNN	Random forest
Accuracy Score	0.5839	0.9998	0.6143	0.6984	0.9998
Mean score of cross validation	0.5842	0.5344	0.5951	0.5585	0.6089
Std of cross validation	0.0112	0.0077	0.0090	0.0077	0.0097

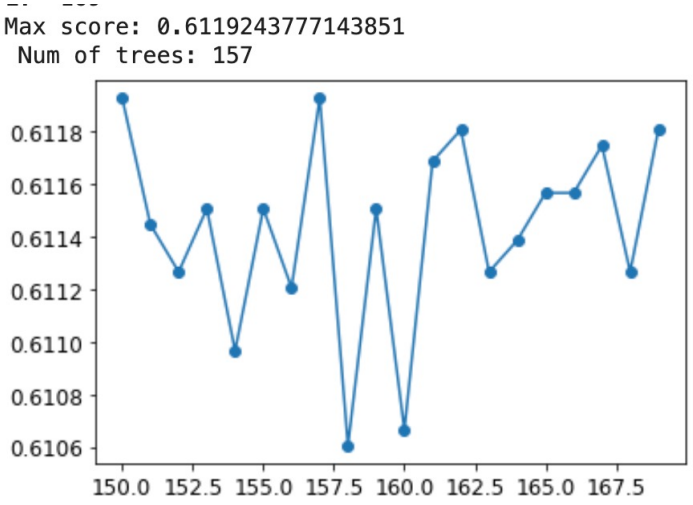
4 categories of causes of fire

	Logistic Regression	Decision tree	SVM	KNN	Random forest
Accuracy Score	0.5476	0.9998	0.5993	0.6955	0.9998
Mean score of cross validation	0.5465	0.5260	0.5837	0.5448	0.6016
Std of cross validation	0.0130	0.0113	0.0110	0.0094	0.0093

Optimal Random Forest

- n_estimators=102, 0.6113 -> 0.6119





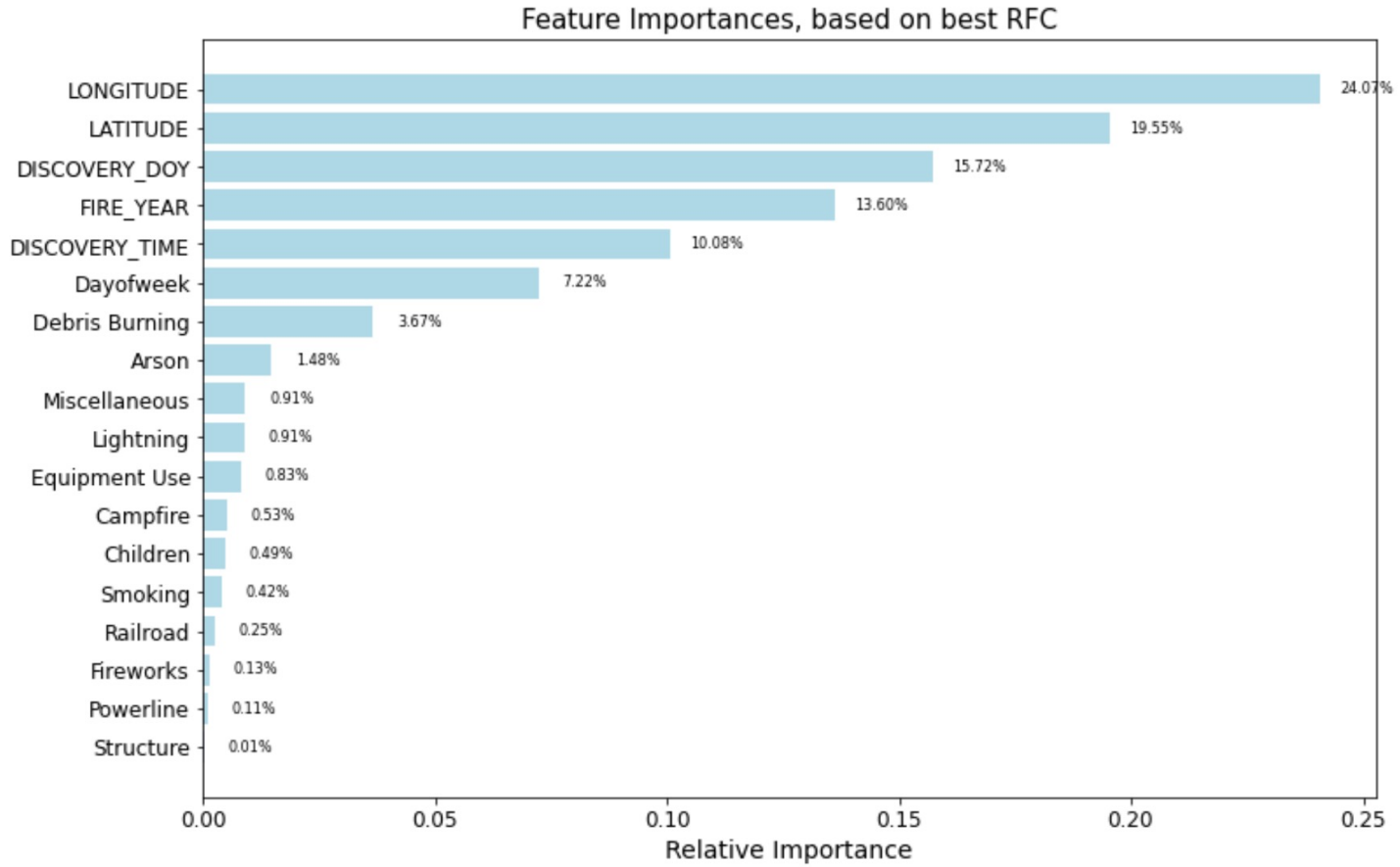
- max_depth=13,
  0.6119 -> 0.6207
- min_samples_leaf=1
  min_samples_split=2,
  0.6207 -> 0.6208
- Criterion='entropy'
  max_features=5,
  0.6208 -> 0.6231

```
GS.best_estimator_
```

```
RandomForestClassifier(criterion='entropy', max_depth=13, max_features=5,
                       n_estimators=157, n_jobs=-1, random_state=90)
```

Apply to the Holdout

- **Predict using testing set, accuracy is 0.6227**
- Feature importance



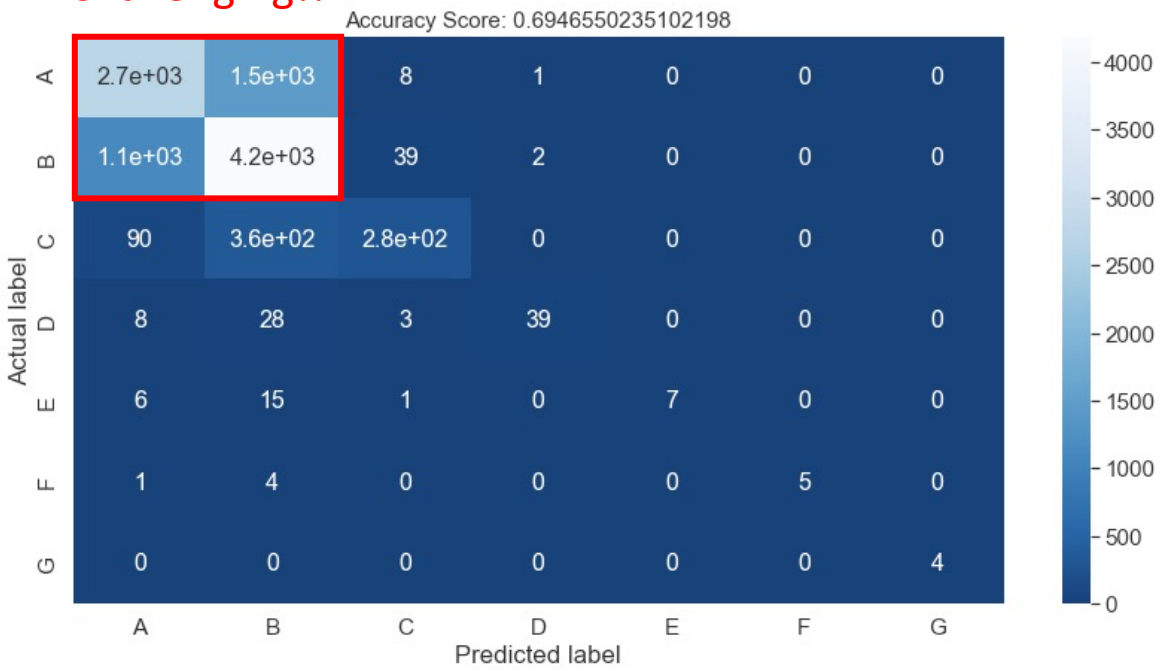
Ensemble Learning with Blending (Base : SVC / MLP / KNN) (Meta : RandomForest)

- **Predict using testing set, accuracy is 0.695**
- **Confusion Matrix**

More important to fire department

	precision	recall	f1-score	support
A	0.68	0.65	0.67	4200
B	0.69	0.78	0.73	5378
C	0.84	0.38	0.53	722
D	0.93	0.50	0.65	78
E	1.00	0.24	0.39	29
F	1.00	0.50	0.67	10
G	1.00	1.00	1.00	4

Challenging!!



Findings from models

Independent single factor is not significantly associated with size of fire based on correlation matrix.

Geographical features and date together are factors that associates with large fire size. We guess all these factors are related to climate, which may be an important factor of fire spread by intuition

If time allows...

- Different angles
  - Predict the causes of wildfires
- Expand the range of study to bigger area.
- Include more related datasets
  - Meteorological data
  - Economic data

Thanks for listening