Wildfire Incident and Property Insurance: A Data-Driven Analysis

AI Studio Final Presentation

Break Through Tech Virtual
December 1, 2023

Introductions

Meet Our Team!

Victoria Nunez
University of Illinois at Urbana-Champaign

Sphia Sadek
University of Pennsylvania

Desarae Cotton
Simmons University

Nisha Shastry
University of Washington

Beruktawit Gebreamlak
Columbia University

Our AI Studio TA and Challenge Advisors

James Donnelly
AI Studio TA - Break Through Tech

Ellie Burns French
Challenge Advisor - Allstate

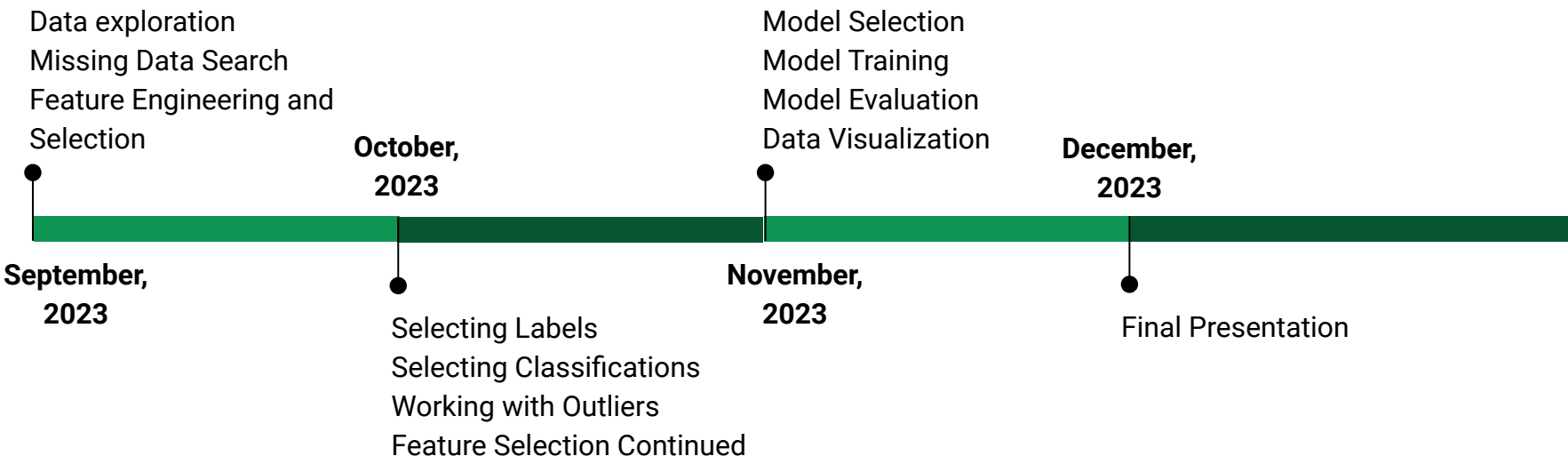
Xieting (Nancy) Zhang
Challenge Advisor - Allstate

Presentation Agenda

1. AI Studio Project Overview
2. Our Approach
3. Data Understanding and Preparation
4. Model Selection and Evaluation
5. Conclusion/Impact
6. Questions

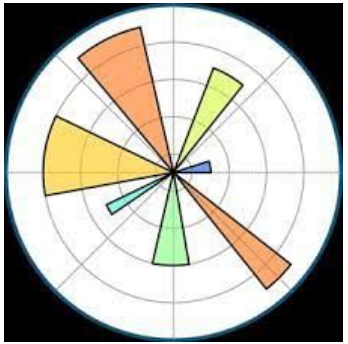
AI Studio Project Overview

Our Model Development Approach



Resources We Leveraged

pandas



scikit learn

seaborn

NumPy

Data Understanding & Data Preparation

Raw Data Set

- Provided by Allstate
- US Wildfire data from 1992-2015
- Data set includes: fire discovery date, cause of fire, etc.
- Lots of metadata

	Unnamed: 0	OBJECTID	FOD_ID	FIRE_YEAR	DISCOVERY_DOY	DISCOVERY_TIME	STAT_CAUSE_CODE	CONT_DOY	CONT_TIME	FIRE_SIZE
count	7.988900e+04	7.988900e+04	7.988900e+04	79889.0	79889.000000	40257.000000	79889.000000	33834.000000	29848.000000	79889.000000
mean	1.315363e+06	1.315364e+06	4.377702e+07	2010.0	187.132659	1456.864396	6.358122	199.431489	1514.818514	43.728864
std	1.803478e+05	1.803478e+05	8.672774e+07	0.0	92.578800	383.144656	3.451852	84.728789	405.815767	1512.638775
min	1.067487e+06	1.067488e+06	1.302561e+06	2010.0	1.000000	0.000000	1.000000	1.000000	0.000000	0.001000
25%	1.195552e+06	1.195553e+06	1.449595e+06	2010.0	100.000000	1250.000000	4.000000	122.000000	1300.000000	0.100000
50%	1.218905e+06	1.218906e+06	1.476542e+06	2010.0	189.000000	1454.000000	5.000000	207.000000	1532.000000	1.000000
75%	1.423229e+06	1.423230e+06	1.950201e+07	2010.0	264.000000	1700.000000	9.000000	266.750000	1800.000000	3.000000
max	1.880429e+06	1.880430e+06	3.003482e+08	2010.0	365.000000	2359.000000	13.000000	365.000000	2359.000000	306113.000000

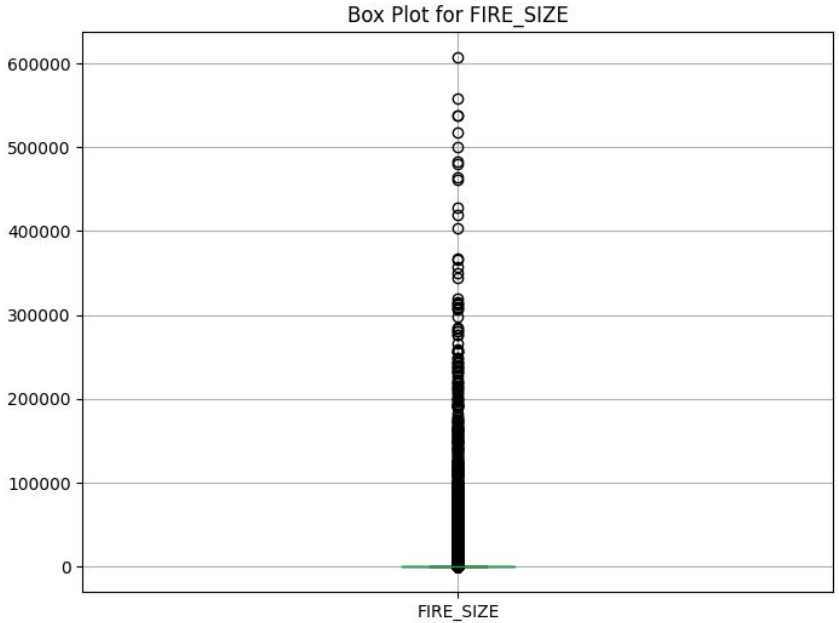
Data Cleaning and Preparation - Outlier Detection (Z-Scores)

Purpose of Outlier Detection:

- Find any abnormal values within fire size
- Not necessary to remove the outliers, they aid our understanding of wildfire data

Findings:

- 3670 outliers in fire size



Data Cleaning and Preparation - Replacing Missing Values

Numerical:

- Few missing values; replaced null with mean

Categorical:

- High volume of null values
- Replaced with the mode value found within entire dataset

Data Cleaning and Preparation - Feature Engineering

What is Feature Engineering:

- “Feature engineering is the process of selecting, manipulating, and transforming data into columns that can be used in machine learning models.”

Label Encoding:

- Transforming categorical features into numerical ones

Reasoning:

- Dataset Size: Due to the scale of the dataset
- RAM Constraints: Google Colab's RAM limitations influenced the decision.
- Efficiency: Minimized columns to prevent training and testing disruptions.

Data Cleaning and Preparation - Feature Engineering

TYPE_FIRE_SIZE_CLASS
3
1
1
0
1
2
1
1
1
0

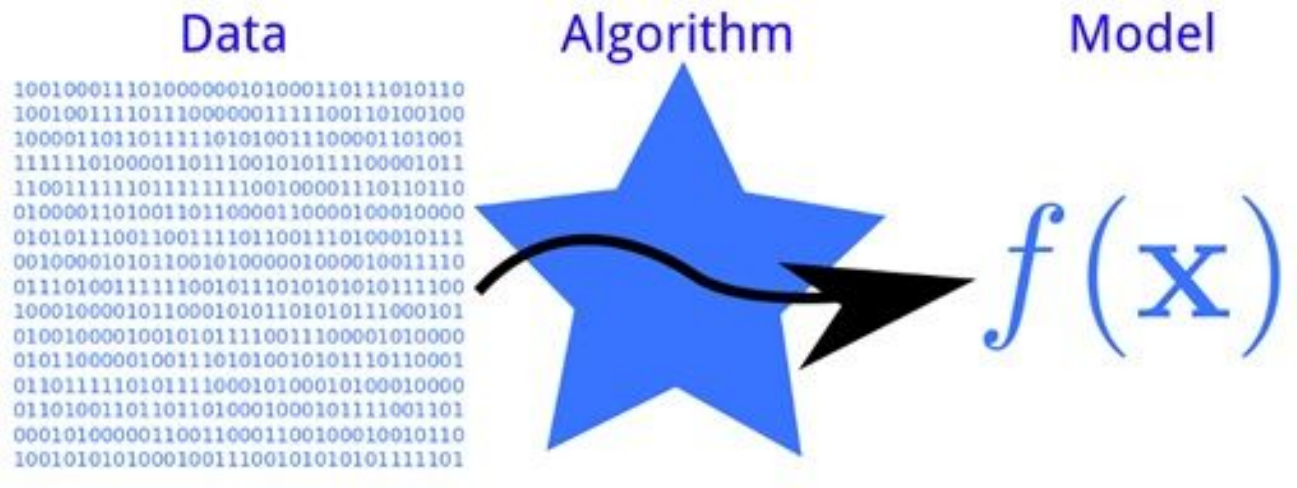
```
df_example_1 = pd.get_dummies(df['FIRE_SIZE_CLASS'], prefix='FIRE_SIZE_CLASS_')
df_example_1
```

	FIRE_SIZE_CLASS__A	FIRE_SIZE_CLASS__B	FIRE_SIZE_CLASS__C	FIRE_SIZE_CLASS__D	FIRE_SIZE_CLASS__E	FIRE_SIZE_CLASS__F	FIRE_SIZE_CLASS__G
0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	1
2	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0
4	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...
74486	1	0	0	0	0	0	0
74487	1	0	0	0	0	0	0
74488	1	0	0	0	0	0	0
74489	0	1	0	0	0	0	0
74490	1	0	0	0	0	0	0

Modeling & Evaluation

Machine Learning Model

This is a program that is **trained** on a dataset to **find patterns** to make **predictions**



Features

These are the **input variables/attributes** in our dataset (column names)

"Allstate_pif_fraction": Allstate's policies in force in the county

"LATITUDE": house's distance north/south of the equator

"LONGITUDE": house's distance east/west of the prime meridian

"Population": the county's population

"Estimated_houses": number of houses in the county

"county_fips": a identification code for the county

"Year": Year a fire occurred in the house

Label

This is the outcome we want our model to predict

Fire Status Cause Code Reference:

1-Lightning Natural
2-Equipment and vehicle use
3-Smoking
4-Campfire Recreation and ceremony
5-Burning Debris and open burning
6-Railroad operations and maintenance
7-Arson
8-Children Misuse of fire by a minor
10-Fireworks
11- Power line Power generation/transmission/ distribution
12-Structure/Firearms and explosive use Undetermined (Human or Natural)

Information on Fire Causes, and how to prevent each one

Instances

These are the examples from our dataset (rows).

Our Label

Year	LATITUDE	LONGITUDE	county_fips	population	estimated_houses	allstate_pif_fraction	STAT_CAUSE_CODE
1994.0	46.715007	-116.334134	16035	8116	2110	0.061611	13.0
2001.0	42.633820	-122.622000	41029	201553	52402	0.086695	1.0
2013.0	40.156667	-120.846667	6063	18710	4863	0.092330	1.0

One instance

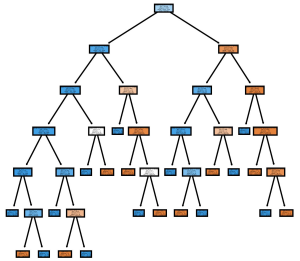
Our Machine Learning Model's Goal

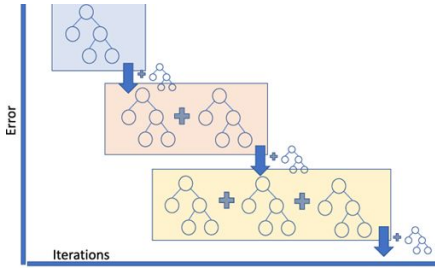
“

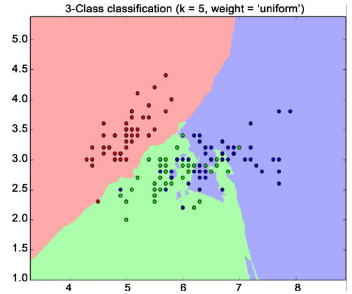
Multi-class Classification Task: Based on county's features, we will predict the most likely fire cause to occur in that geographical location.

Model Selection Process

- **Features**
  - "FIRE_YEAR", "population","LATITUDE","LONGITUDE", "estimated_houses", "allstate_pif_fraction","county_fips"
- **Label**
  - The code of the statistical cause of the fire.
- **Chosen Algorithms**
  - Decision Trees/Random Forest
  - KNN
  - Gradient Boosting







Understanding Fire Causes: Our Model's Business Impact

- **Human Influence**: Nearly **85%** of wildland fires are caused by human activities.*
- **Natural Causes**: The remaining **15%** are predominantly due to lightning strikes.

*Source: 2000-2017 data based on Wildland Fire Management Information (WFMI) and U.S. Forest Service Research Data Archive

Photo Credit: National Park Service

Understanding Fire Causes: Our Model's Business Impact

Understanding the causes of fires in specific areas enables AllState to:

- **Assess Risk**: If the cause is natural (and thus unpreventable), recognize that the area poses a higher risk to the company.
- **Community Engagement**: If the cause is human (and thus preventable), inform community members and policyholders about actions and steps they can take to help reduce the risk.
- **Policy Adjustment**: Modify insurance policies based on the predominant cause of fires in the area.

Our Model's Business Impact

Wildfires increase the risk of flooding

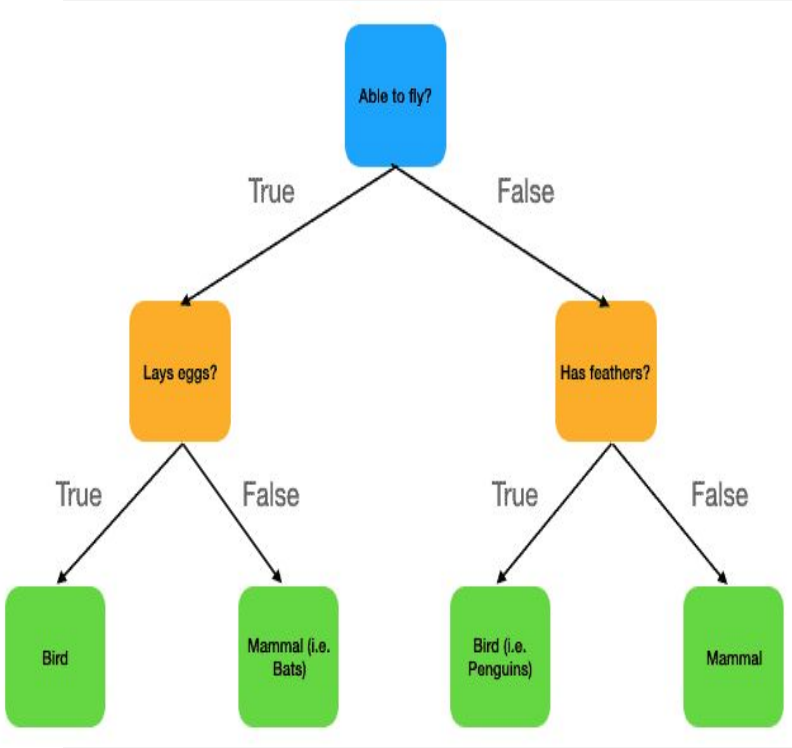
"Burn scars leave the ground very dry and unable to absorb water, and the terrain can even become as dense as pavement." -*Allstate Flood insurance Resources*

Thus Understanding Fire Causes can positively impact the flood insurance sector

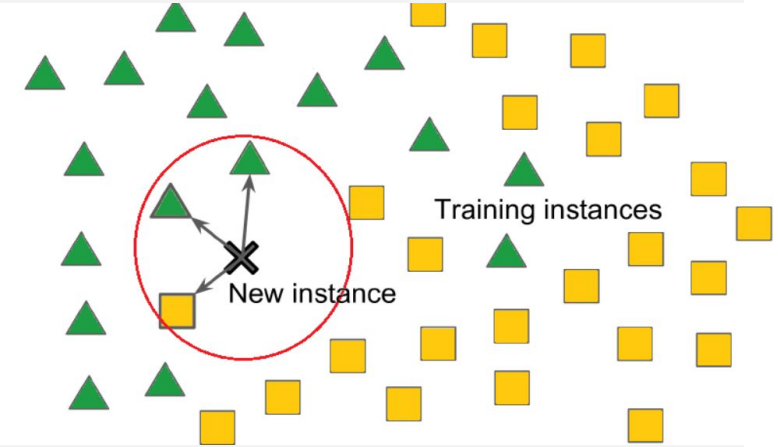
Reed Saxon | Credit: AP (Nov. 29, 2018)

Decision Trees + Random Forest

- Follows a flowchart-like structure, every branch in the tree is a decision-making process
- The leaves are the ultimate outcome
- For classification problems, the prediction is based upon the majority class
- For regression problems, the prediction is based on the average in the leaf node
- Random forest is an ensemble method that combines multiple of these decision trees to make decisions



KNN (K-Nearest Neighbors)



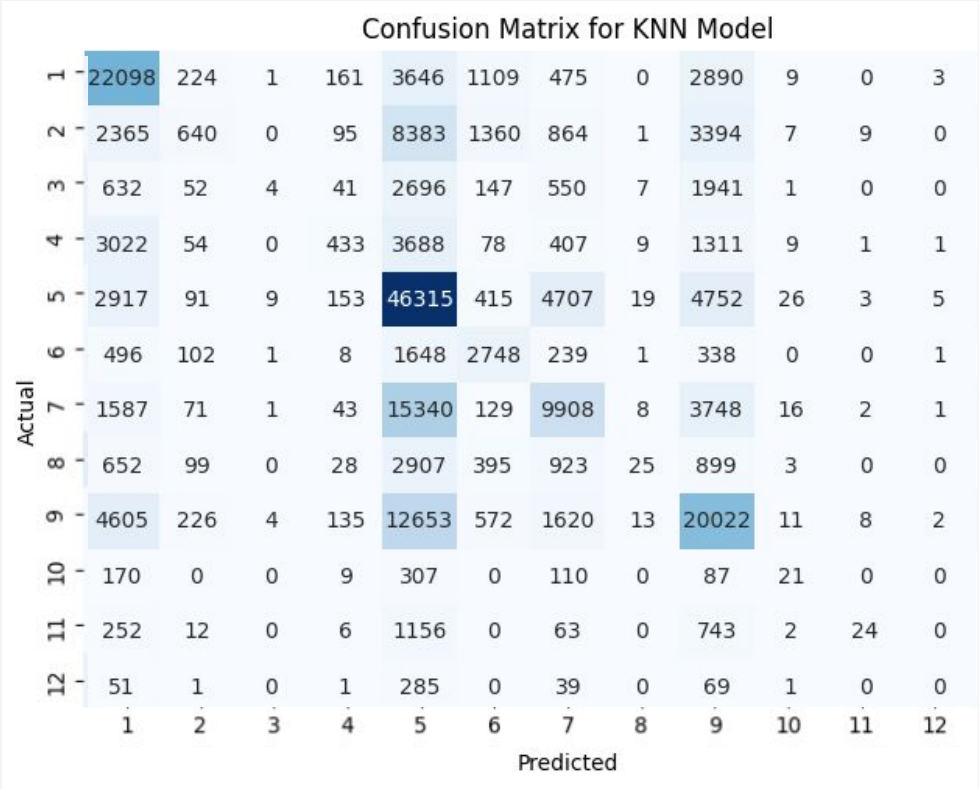
- KNN is a model method that utilizes neighboring data points to make predictions about a new point
- K is the number of “neighbors” you consult to help make a prediction
- The features of the majority of the nearby points determine what features the new point will be predicted to have

Model Comparison

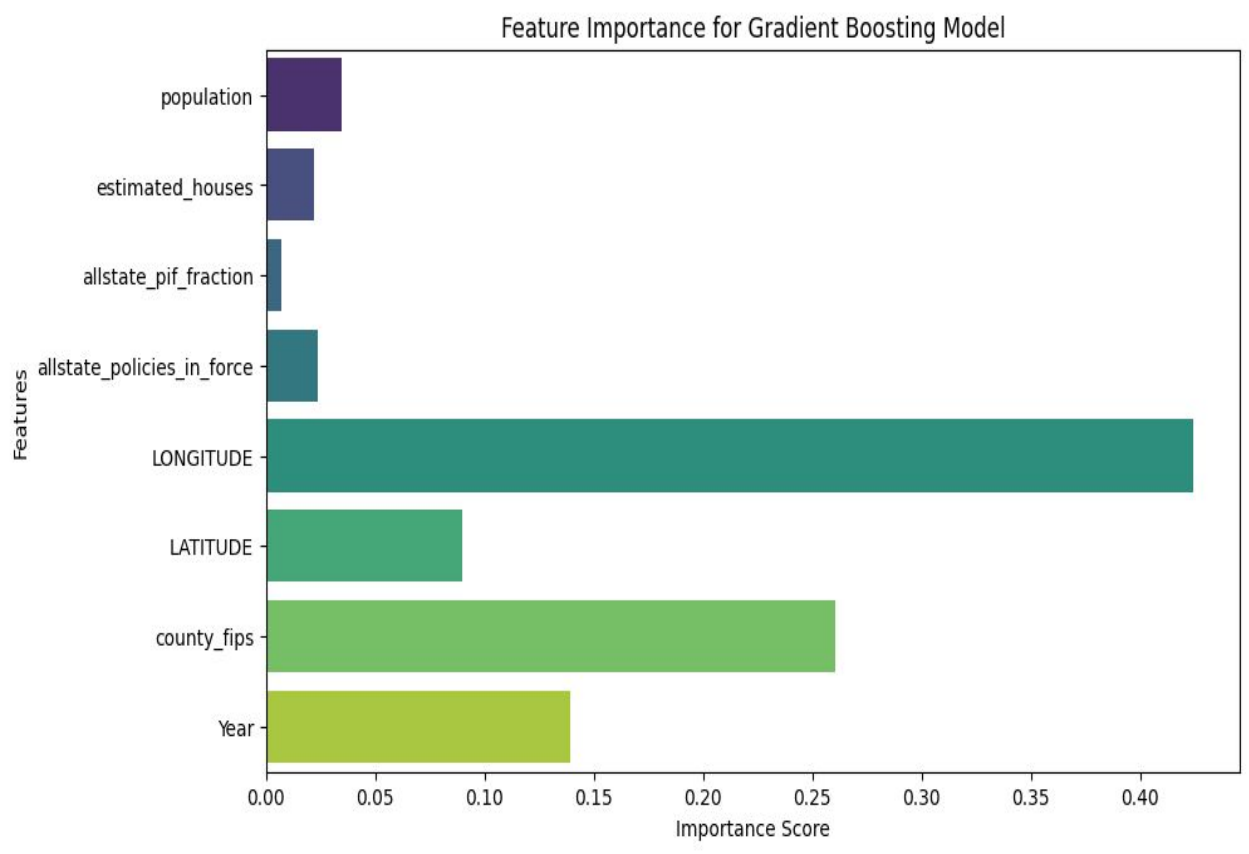
Model Name	Description	Accuracy Results	Pros	Cons
Decision Trees	Tree structure where the leaves are the final decisions and the branches are decisions during the process of making predictions	54%	A lot more intuitive and easier to interpret	Harder to pick up nuances in data therefore accuracy scores are not great
Random Forest	enhances accuracy by combining predictions from an ensemble of decision trees.	58%	Prevent overfitting in training data. Work well with regression and classification models	For larger datasets, random forest can become slow to evaluate, as the trees become too large
KNN	Supervised machine learning model for classification and regression tasks, makes predictions by calculating distances between points	56%	High adaptability, and works well with both regression and classification models	Becomes computationally expensive over time to calculate the distances between each data point with large data sets
Gradient Boosting	enhances accuracy by calculating residuals (errors) of previous models and learning from those errors	51%	Prevents overfitting in training data, and useful for complex data or non-linear patterns	Oftentime requires hyperparameter tuning alongside the model and can be sensitive to outliers

Model Visualization (KNN)

- The diagonal values in the confusion matrix are the # of instances correctly predicted
- Ideally everything outside the diagonal values would be 0, those represent misclassifications
- Class 5.0 (Burning Debris) was best wildfire predicted, it's the darkest color in our confusion matrix with 46,315 accurate predictions
- The classes represent one of the categories of 'STAT_CAUSE_CODE' or the fire type



Model Visualization (Gradient Boosting)



- Helps us view which feature has the most impact on our gradient boosting model
- The features with higher importance score contribute to more of the model's predictions
- Longitude is the most influential for our gradient boosting model, meaning longitude or east/west geographical location impacts wildfire cause predicted the most

Model Visualization (Gradient Boosting)

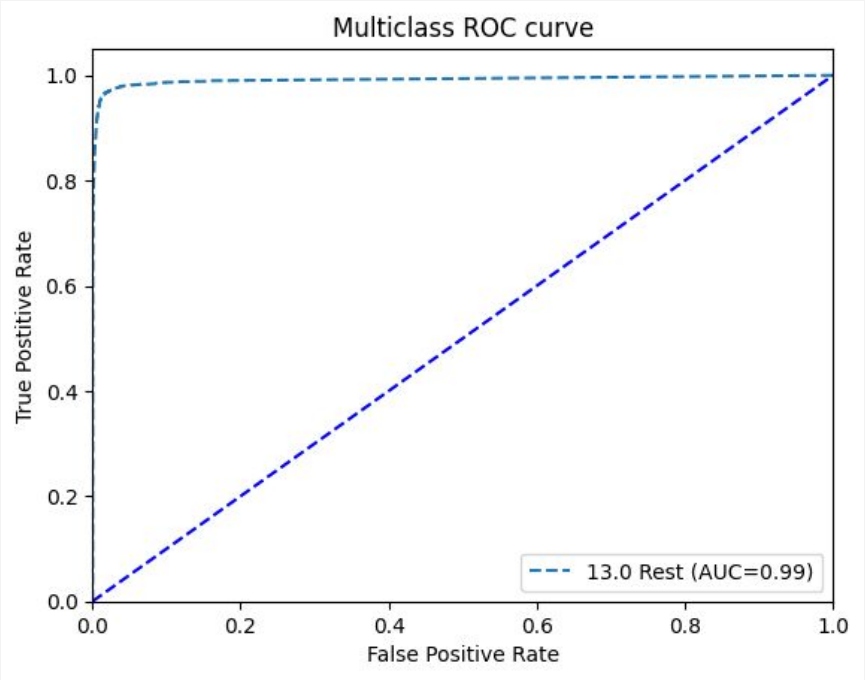
Classification Report:

	precision	recall	f1-score	support
1.0	0.53	0.71	0.61	31041
2.0	0.40	0.04	0.07	17852
3.0	0.17	0.00	0.00	6423
4.0	0.38	0.05	0.08	9271
5.0	0.46	0.75	0.57	61379
6.0	0.40	0.49	0.44	5637
7.0	0.50	0.30	0.38	32684
8.0	0.30	0.00	0.01	6340
9.0	0.48	0.49	0.49	41234
10.0	0.19	0.03	0.05	733
11.0	0.50	0.01	0.02	2294
12.0	0.00	0.00	0.00	448
13.0	0.73	0.80	0.76	24950
accuracy			0.51	240286
macro avg	0.39	0.28	0.27	240286
weighted avg	0.48	0.51	0.46	240286

- Model predicts certain classes really well (ex. class 1.0 which is lightning natural), while others are not (class 10.0 which is fireworks).
- Precision is the accuracy of positive predictions
- Recall is the percentage of instances the model correctly identifies a class
- f1-score is the mean of precision and recall
- Support is the number of occurrences in the dataset

Model Visualization (Decision Trees/Random Forest)

- AUC (Area Under the Curve) score is 0.99 which is normally good, the model distinguishes different classes
- Low-false positive rate, which means the model is not misclassifying instances
- However, the AUC is often influenced by the dominant class or in our case, the wildfire with the most instances (campfires for example)
- This means a class imbalance could be the cause of our high AUC score, so we need to use different metrics to evaluate our decision tree model (f1, precision, recall, etc.)



Analysis

- **Accuracy is not the best assessment metric for multi-classification problems,** might disregard the minority classes
- Accuracy is measured correct predictions / all predictions, this level of accuracy is blind to specific classes
- Calculating **precision and recall** for each class individually is helpful for assessing a multi-class classification model
- When using direct accuracy scores for each of our model, it seems they perform poorly in terms of accuracy which may be due to blindness to specific classes having **less instances than others**!
- Our **random forest model (ensemble technique using several decision trees) was most accurate** in determining the most likely fire cause to occur in a geographical location, however there is still work to do to make these models **even more accurate**!

Case Study (Lightning Caused Fires)

A surprising amount of dry lightning hits California, fueling fire risk

Lightning-caused fires are more prevalent in the northern half of the state, particularly over mountainous terrain

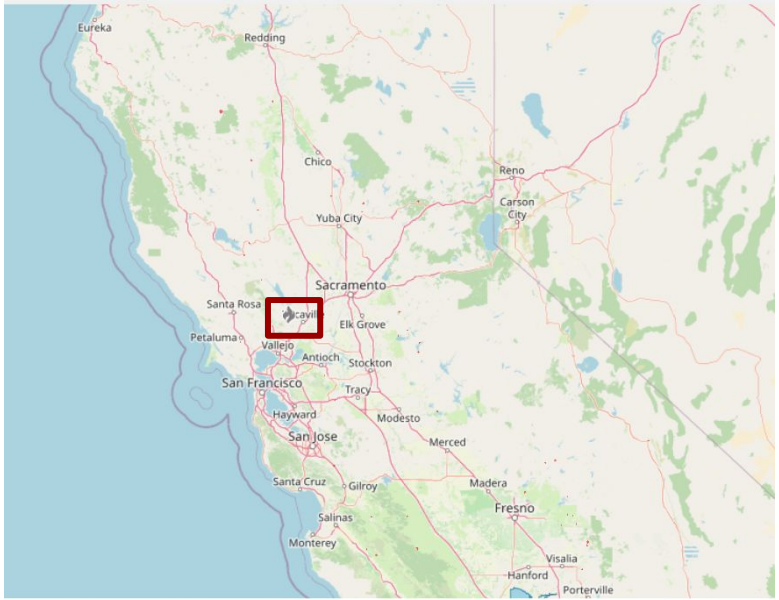
By Mike Branom

August 24, 2022 at 10:48 a.m. EDT

Lightning strikes over Mount St. Helena in 2021 in Napa County, Calif. Northern California was hit by more than 1,000 strikes during the storm. (Kent Porter/AP)

Case Study (LNU Lightning Complex fires 2020)

(AP Photo/Noah Berger)



Map data © OpenStreetMap contributors, CC-BY-SA

Case Study (LNU Lightning Complex fires 2020)

(Active for 46 days)

1,491 structures destroyed, **232 damaged**

Affected Counties: *Sonoma, Napa, Yolo, Solano, and Lake Counties*.

5th deadliest wildfire in California History

How it began?
From several small lightning-sparked fires that merged together.

Statistics taken from Frontline Wildfire Defense

Case Study (LNU Lightning Complex fires 2020)

Using Random Forest Model to testing on an unseen instance

"county_fips": 6055

"population": 130000

"Estimated_houses": 33800

"Allstate_pif_fraction": 7.23%

“LATITUDE": 38.48193

"LONGITUDE": -122.14864

"Year": 2020

Case Study (LNU Lightning Complex fires 2020)

Results

"county_fips": 6055

"population": 130000

"Estimated_houses": 33800

"Allstate_pif_fraction": 7.23%

“LATITUDE": 38.48193

"LONGITUDE": -122.14864

"Year": 2020

Model Predicted Fire Cause Code: 1.0

Final Thoughts

What We Learned

- A multi-classification model comes with several challenges:
- Imbalanced classes, which lead to biased models that are able to have good performance on the majority but poor performance on the minority
- The model is sensitive to outliers and missing data, so more time should be sent on preprocessing
- When working with a large data set, it's important to consider optimization techniques for certain models, some of our models took hours to run!
- Evaluation metrics for a machine learning model varies based on the type of model, for ours, accuracy was not always the most useful! (ex. precision, recall, f1 scores, confusion matrix analysis, etc.)

Potential Next Steps

- Make our models have a higher prediction accuracy! We averaged around ~55% accuracy per model, which can be improved with more time.
- Add more steps into our data preprocessing steps, and determining the pros/cons of removing NaN values or using one-hot encoding on categorical data
- Explore more of the correlations between different features within each model we created, some models have features that are highly correlated with one another which can skew predictions!
- Create forecasting models that use algorithms that are suited for our temporal data


Questions?