Final Project Report

William Huang

Data Overview

2.2 Describe data

Two datasets are provided for this assignment. The first dataset contains information on historical fires that occurred in California through 2020. These fires were all wildfires, meaning that they were not made intentionally by humans. The dataset includes information such as the year of the fire, the fire name, and the geographical location of the fire. There are 20,772 observations in this dataset. The dataset contains 19 total columns, where one column is an id column identifying each specific fire. The other 18 columns represent variables that describe the fires.

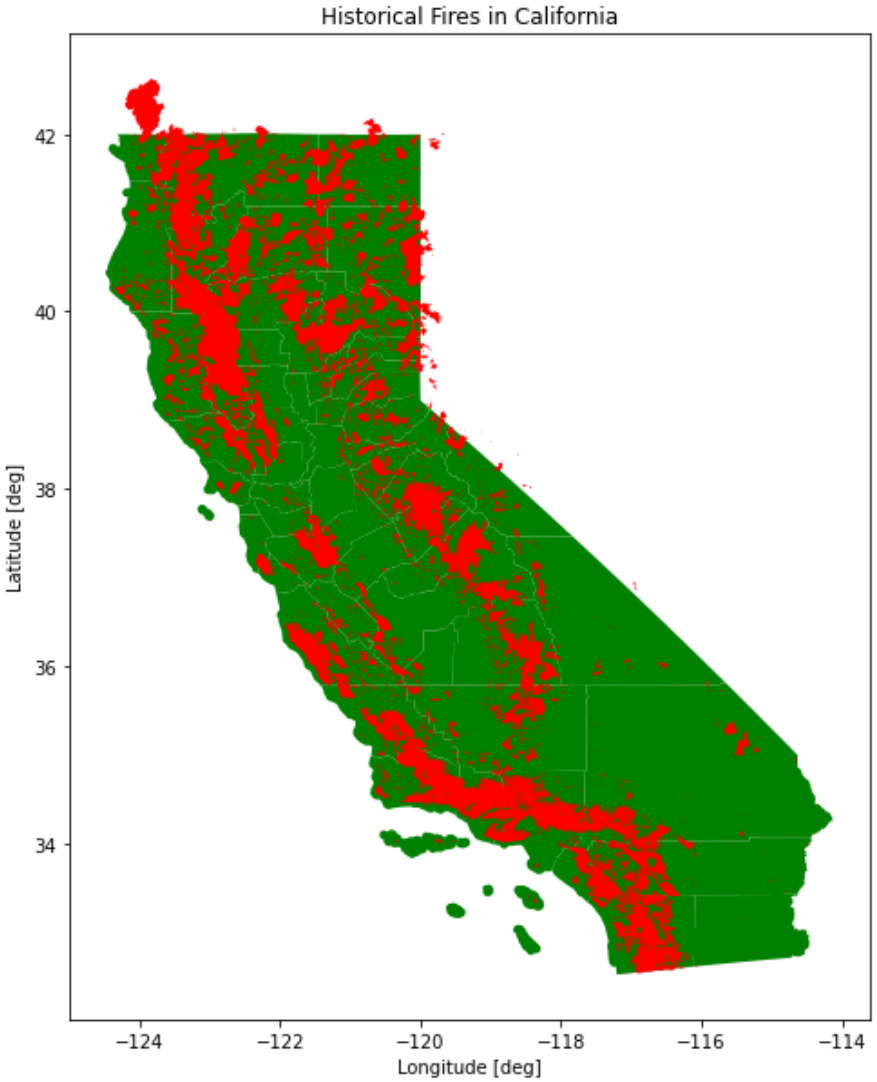
The second dataset (attributes dataset) that was provided contains wildfire-related information on 5 km grids across the earth in California. Some variables included in this dataset are temperature, precipitation, and wind speed. There are 7,338,672 observations in this dataset. The dataset contains 19 total columns, where two columns represent the unique latitude and longitude of the 5 km grid. The other 17 columns represent variables that describe the grids of land.

An additional dataset was used that contains information on the counties of California and their geographical location. There were only two columns in this dataset, and there were 58 rows that represented each county in California.

The data acquired did satisfy the requirements, as these datasets provide enough information to predict whether a fire will occur on these 5 km grids in California.

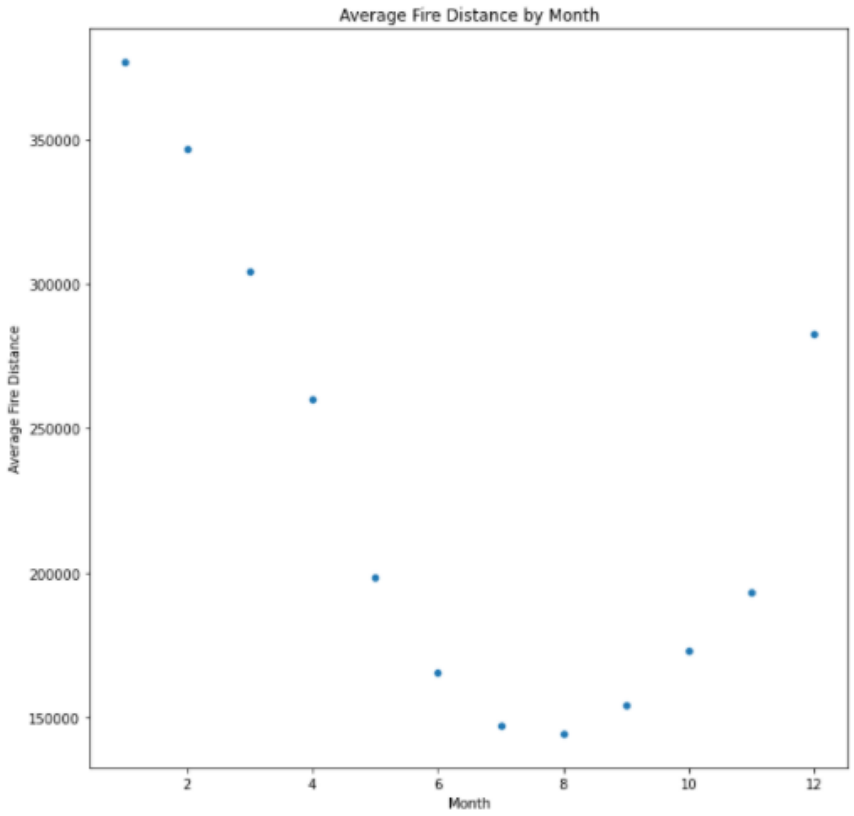
2.3 Explore data

To start, I was first interested in viewing where the historical fires in California typically occur.



The visualization above shows the fires (red) that have occurred in California through 2020. As we can see, fire is highly dependent on location. Much of the fire occurs along the coast, where temperature and wind is much different compared to the inland of California. Therefore, this visualization illustrates that the wildfire-related variables in the attribute dataset are important in the prediction of fires.

In addition, I wanted to determine how important the months are for when fires occur. This will decide if I will keep each 5 km grid separated by month or not.

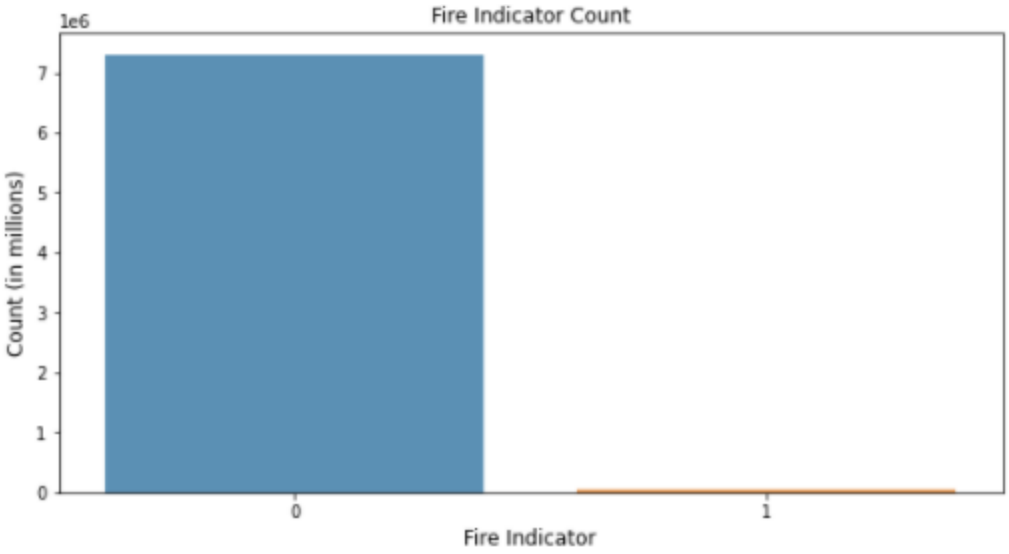


The visualization above shows the average fire distance by month for the 5 km grids. The fire distance variable determines the meters separating the 5 km grids and the nearest occurring fire during that month. As we can observe, the months during warmer temperatures (spring and summer) have lower average fire distances compared to months during colder weather (fall and winter). This makes sense as more fires would occur during higher temperatures, meaning that these 5 km grids have a greater chance of being closer to these fires. Therefore, this visualization reveals that the month variable is important in determining when fires occur.

2.4 Verify data quality

Data quality

A quality issue found is how imbalanced the dataset is. As the goal is to predict if a wildfire occurs in the 5 km grids, a new variable was created called “fire”. This is a simple target variable that is 0 if the fire distance variable is greater than 5000 meters (outside of the 5 km grid) or a null value. If the fire distance is less than or equal to 5000 meters, the target fire variable is 1. Below shows a bar plot of the distribution for the target variable.



As we can see, there are significantly more 0s in the target variable than 1, which would harm our ability to predict effectively. Imbalance in a dataset leads models to simply predict the target value that heavily outweighs the other value, so this issue will need to be addressed before model creation. One solution is resampling without replacement and with weights, which would lead to the target column being more balanced.

In addition, there were null values that were present throughout the numeric columns. One solution is to replace the null values with the median of the entire numeric column.

Data Preparation

3.1 Data selection

The first dataset was excluded from the final dataset. This is because it was extremely difficult to find a common variable to join this dataset with the attributes dataset, and the attributes data frame contains extremely important wildfire variables that are needed to predict fires. One way to incorporate the first dataset's information is to create a number of historical fires columns in the attributes dataset. This can be done by checking if the points for the 5 km grids were in the geometric area of the historical fires. However, both datasets had an immense amount of data, so this was infeasible to accomplish in a reasonable amount of time.

Therefore, the attributes dataset was chosen to be included in the final dataset as it had more important variables for fire prediction compared to the first dataset. The dataset on California counties was also included as this would be important information in determining if a fire were to occur or not for these 5 km grids. All the variables were kept in the attributes dataset besides geometric location variables and the fire distance variable. The geometric data contained variables related to latitude and longitude, which were not included because this would be

redundant location information that was already present in the county data. The first distance variable was dropped as a new fire indicator variable was created from this column.

3.2 Data cleaning

To address the imbalanced data issue, I resampled the data with independence. Each observation had weights (1s had much more weight than 0s), which balanced out the scare observations with fire in their grid with the observations without fire. This fixed this issue of imbalance, and the final dataset contained a much more even distribution of our target column.

To address the null values in the numeric columns, the null values were replaced with the median value for the entire numeric column.

3.3 Feature engineering

Derived attributes

The “fire” indicator variable was created. This is a simple target variable that is 0 if the fire distance variable is greater than 5000 meters (outside of the 5 km grid) or a null value. If the fire distance is less than or equal to 5000 meters, the target fire variable is 1.

3.4 Data integration

As stated under the data selection section, the attributes dataset provided was used along with a dataset containing information on california counties and their geographical location. The attributes dataset was inner joined spatially with the counties dataset. Simply put, the attributes dataset gained a new column that specified which county the 5 km grids were in. Any row that did not have a California county was dropped from the final dataset due to the nature of inner join.

3.5 Data formatting

Standard scaling was implemented on all the numerical features in the dataset because it helps with the accuracy of many classification models.

Data Modeling

4.1 Modeling technique

The two modeling techniques used for binary classification are Logistic Regression and Random Forest.

4.2 Test design

Describe the plan for training, testing, and evaluating the models.

The final dataset was split into X, which contains all the predictor columns, and y, which contains the target variable. The X and y were then split into 70% training data and 30% testing data. The three models will be trained on the training data and will be evaluated based on their predictive accuracy as well as their f1 score.

4.3 Model creation

Models produced

Linear Regression:
Intercept: -0.735

10 most important variables + coefficients: (95 total coefficients)

	variable	coefficient
12	vpd	-22.985741
3	ppt	9.669540
4	q	9.660511
7	swe	9.659112
8	tmax	4.801913
10	vap	-4.741860
13	PDSI	-4.509695
1	def	4.238047
9	tmin	-4.232303
0	aet	3.345246

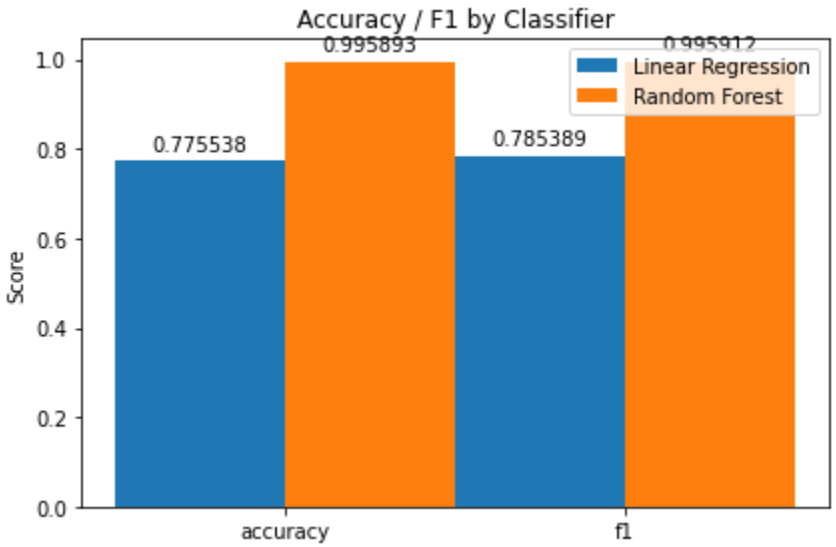
Note: *Appendix* contains full coefficients of the model.

Random Forest:

```
rfc = RandomForestClassifier(n_estimators = 10)
```

Note: Random Forest uses decision trees that cannot be displayed visually.

4.4 Model assessment



The two metrics used to evaluate our models as accuracy and f1 score. From the chart above we can clearly see that random forest has a much higher accuracy and f1 score. While the linear regression model has an accuracy of 0.776 and a f1 score of 0.785, the random forest model has an accuracy of 0.996 and f1 score of 0.996. Therefore, the random forest model will be the final model that we will use for prediction.

Evaluation

5.1 Results evaluation

In terms of the project success, the criteria was met. A model was created that was able to predict wildfires at 5 km resolution at an accuracy of 99.6% as well as have an f1 score of 99.6%. Therefore, the objective of using machine learning to predict California wildfires was completed based on the results of the final model.

5.2 Discussion

One thing that I found to be concerning was integrating the attributes dataset with the county dataset. After inner joining the two datasets, around half of the attribute dataset's observations

were lost, which would only happen if some of the 5 km grids in the attributes dataset were not in any of the 58 California counties. However, the attributes dataset was gathered from data on California from 2010-2020, so this problem should not have occurred. If I had more time, I would look more closely into this issue.

This analysis could be misleading as people could assume that this model could be used to predict wildfires in general. However, this model can only be used to predict wildfires in California, as this was the only location that the data was gathered from. Additional data from other areas would be needed for the final model to apply more broadly.

One misinterpretation of the results is that the model can predict wildfires based on county. The model can only predict based on the 5 km grids in California, so if the model predicts no fire for the grid, this does not mean that a fire will not occur relatively close to this area. In addition, the model will never 100% accurately predict wildfires, regardless of the amount of data given to it. Therefore, we must inform users that they should take the model with a grain of salt and not fully believe its conclusions, as there is always a possibility of error.

Appendix

Logistic Regression Intercept and all 90 Coefficients:

```
(array([-0.73477404]),
 array([[ 3.34524561e+00,  4.23804688e+00, -1.71518474e+00,
          9.66954040e+00,  9.66051089e+00, -4.38971088e-01,
          3.11044121e-01,  9.65911168e+00,  4.80191338e+00,
         -4.23230307e+00, -4.74186017e+00, -2.51953217e+00,
         -2.29857409e+01, -4.50969455e+00, -3.05209931e-01,
         -1.09017351e+00, -1.74597472e-01,  1.01709917e+00,
          4.41715313e-01, -3.36534531e-01, -6.10694928e-01,
          5.04477552e-01,  2.48580703e-01,  2.30143710e-01,
         -1.33355561e-01,  9.38674406e-02, -2.22708977e+00,
         -2.26788760e+00,  2.21833744e-01, -1.27756984e+00,
          6.28661808e-01, -8.43565090e-01,  1.26388360e+00,
          1.50155064e-01,  5.78565707e-01,  1.26918135e+00,
         -3.60165150e-02, -1.12130705e+00, -6.41027766e-01,
         -1.11830529e+00,  2.56439096e-01,  9.83146821e-01,
          7.17194149e-02,  1.38178498e+00,  9.88083316e-01,
          1.87709442e-01,  1.65664661e-03, -7.17573972e-01,
         -7.20456494e-01, -1.49366479e+00,  1.68661555e-01,
          3.66053812e-01, -1.04238659e+00,  1.62910206e-01,
         -3.82940698e-01,  5.71357683e-01, -3.94839417e-01,
         -1.34835332e-01,  2.12639314e-01, -1.11993620e+00,
          4.79603990e-01, -1.93906802e-01,  4.78870601e-01,
         -2.46561358e-01, -1.90622680e-01,  4.36064153e-01,
          4.35802691e-01,  5.05220593e-01,  7.50088218e-01,
          1.75269666e+00, -1.51962780e-01,  1.48728545e+00,
         -1.75103029e+00, -2.03380785e+00, -1.84299754e+00,
         -1.17307442e+00,  6.28527541e-02,  1.11891953e+00,
          1.75908019e+00,  1.85912570e+00,  1.49333514e+00,
          6.63623437e-01,  8.58666256e-02, -8.88955426e-01,
         -7.51884885e-01, -1.51810184e-01, -2.16758279e-01,
         -4.31103329e-01, -8.87622835e-01, -3.68924590e-01,
         -5.74945831e-02,  7.60190591e-01,  2.64003497e-01,
          2.30329949e-01,  9.64012487e-01]]))
```