Data 301 - Predicting Fire Causes

Edward Du, Ethan Choi, Cal Schwefler

Why wildfires?

- Regular wildfires ravage the Pacific Coast of the continental United States
- These fires are started for a number of reasons and force many out of their homes
- Our model could help predict the size and duration of fires in the contiguous United States based on time of year
- Heatmap suggests much larger fires in Western half of U.S.

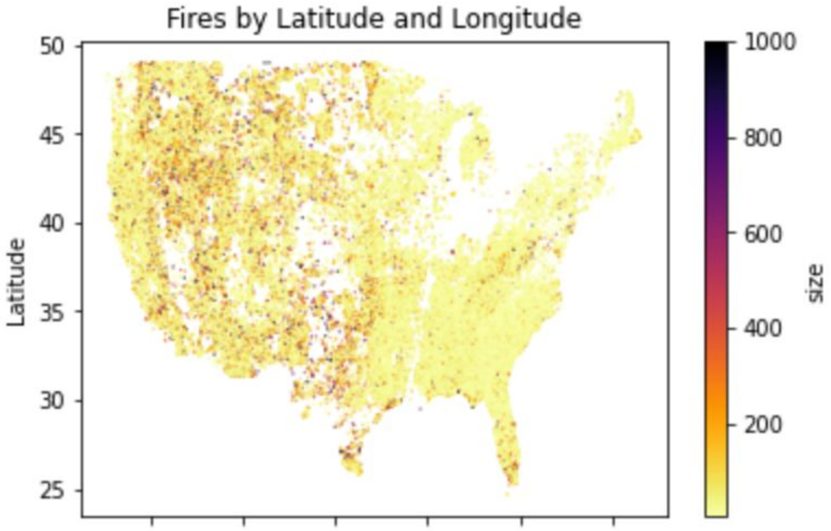
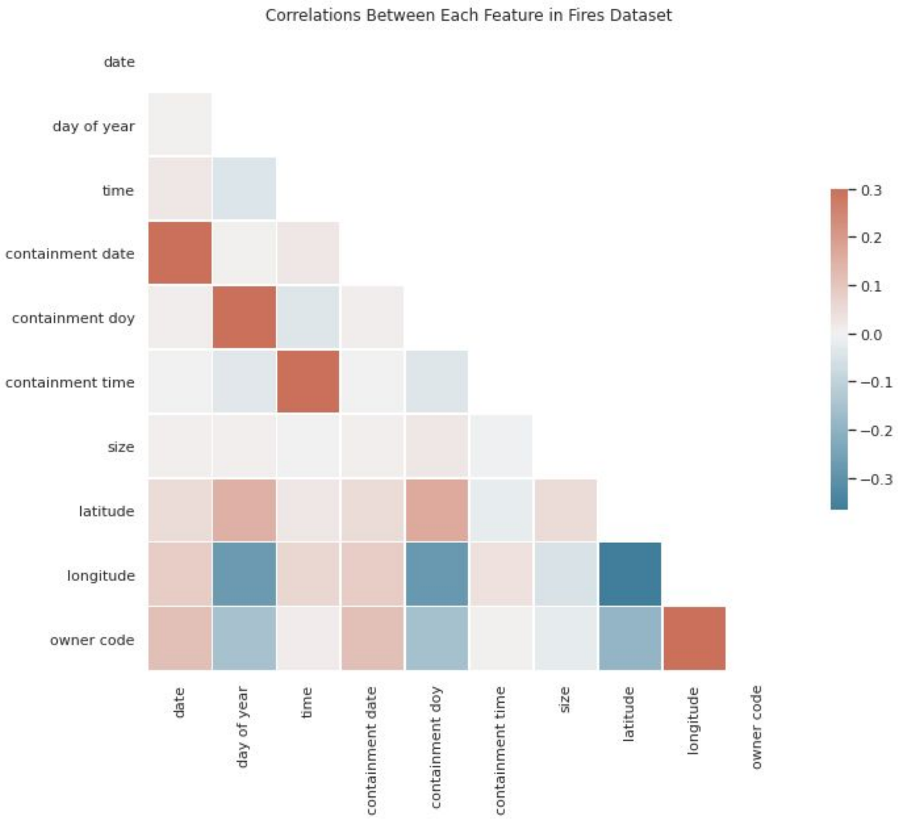


Figure 1: Heatmap of fires from data set. Limited to contiguous 48 states by latitude/longitude and color coded by fire size

Overview of dataset

- Downloaded from Kaggle
- Contains 1.88 million geo-referenced wildfires that occurred in the US from 1992 to 2015.
- Downloaded and read as an sqlite file

```
import sqlite3
import numpy as np
import pandas as pd

pd.set_option('display.max_columns', None)

conn = sqlite3.connect('/content/drive/My Drive/Cal Poly/DATA 301/DATA 301 Project/FPA_FOD_20170508.sqlite')

df = pd.read_sql_query("SELECT * FROM Fires", conn)

print(df)


```



Contents of dataset

```
         OBJECTID     FOD_ID              FPA_ID SOURCE_SYSTEM_TYPE  \
0               1          1          FS-1418826                FED
1               2          2          FS-1418827                FED
2               3          3          FS-1418835                FED
3               4          4          FS-1418845                FED
4               5          5          FS-1418847                FED
...           ...        ...                 ...                ...
1880460   1880461  300348363  2015CAIRS29019636             NONFED
1880461   1880462  300348373  2015CAIRS29217935             NONFED
1880462   1880463  300348375  2015CAIRS28364460             NONFED
1880463   1880464  300348377  2015CAIRS29218079             NONFED
1880464   1880465  300348399  2015CAIRS26733926             NONFED

        SOURCE_SYSTEM NWCG_REPORTING_AGENCY NWCG_REPORTING_UNIT_ID  \
0         FS-FIRESTAT                    FS                USCAPNF
1         FS-FIRESTAT                    FS                USCAENF
2         FS-FIRESTAT                    FS                USCAENF
3         FS-FIRESTAT                    FS                USCAENF
4         FS-FIRESTAT                    FS                USCAENF
...               ...                   ...                    ...
1880460      ST-CACDF                 ST/C&L                USCASHU
1880461      ST-CACDF                 ST/C&L                USCATCU
1880462      ST-CACDF                 ST/C&L                USCATCU
1880463      ST-CACDF                 ST/C&L                USCATCU
1880464      ST-CACDF                 ST/C&L                USCABDU

         NWCG_REPORTING_UNIT_NAME SOURCE_REPORTING_UNIT  \
0          Plumas National Forest                  0511
1        Eldorado National Forest                  0503
2        Eldorado National Forest                  0503
3        Eldorado National Forest                  0503
4        Eldorado National Forest                  0503
...                           ...                   ...
1880460       Shasta-Trinity Unit                 CASHU
1880461   Tuolumne-Calaveras Unit                 CATCU
1880462   Tuolumne-Calaveras Unit                 CATCU
1880463   Tuolumne-Calaveras Unit                 CATCU
1880464       San Bernardino Unit                 CABDU

        SOURCE_REPORTING_UNIT_NAME LOCAL_FIRE_REPORT_ID LOCAL_INCIDENT_ID  \
0           Plumas National Forest                    1            PNF-47
1         Eldorado National Forest                   13                13
2         Eldorado National Forest                   27               021
3         Eldorado National Forest                   43                 6
4         Eldorado National Forest                   44                 7
...                            ...                  ...               ...
1880460        Shasta-Trinity Unit               591814            009371
1880461    Tuolumne-Calaveras Unit               569419            000366
1880462    Tuolumne-Calaveras Unit               574245            000158
1880463    Tuolumne-Calaveras Unit               570462            000380
1880464  CDF - San Bernardino Unit               535436            003225

        FIRE_CODE                FIRE_NAME ICS_209_INCIDENT_NUMBER  \
0            BJ8K                 FOUNTAIN                    None
1            AAC0                   PIGEON                    None
2            A32W                    SLACK                    None
3            None                     DEER                    None
4            None                 STEVENOT                    None
...           ...                      ...                     ...
1880460      None                 ODESSA 2                    None
1880461      None                     None                    None
1880462      None                     None                    None
1880463      None                     None                    None
1880464      None  BARKER BL  BIG_BEAR_LAKE_                    None

        ICS_209_NAME MTBS_ID MTBS_FIRE_NAME COMPLEX_NAME  FIRE_YEAR  \
0               None    None           None         None       2005
1               None    None           None         None       2004
2               None    None           None         None       2004
3               None    None           None         None       2004
4               None    None           None         None       2004
...              ...     ...            ...          ...        ...
1880460         None    None           None         None       2015
1880461         None    None           None         None       2015
1880462         None    None           None         None       2015
1880463         None    None           None         None       2015
1880464         None    None           None         None       2015
```

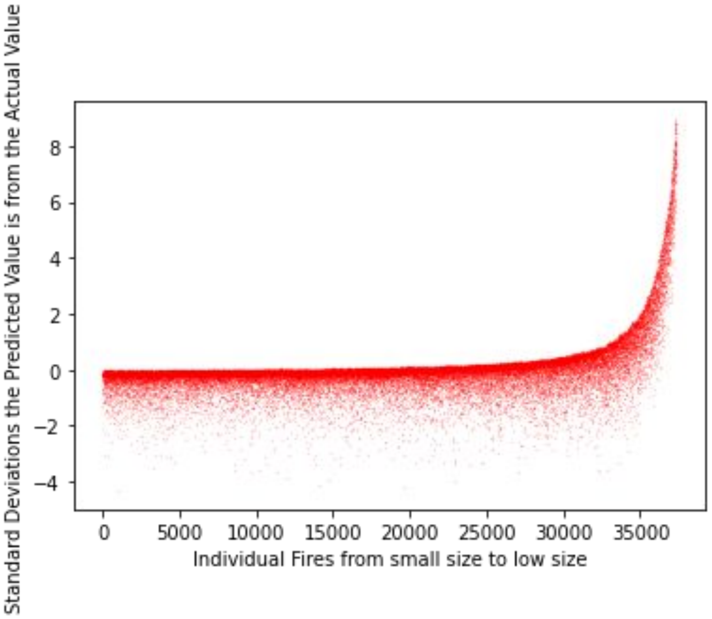
Contents of dataset

```
         DISCOVERY_DATE  DISCOVERY_DOY DISCOVERY_TIME  STAT_CAUSE_CODE  \
0             2453403.5             33           1300              9.0
1             2453137.5            133           0845              1.0
2             2453156.5            152           1921              5.0
3             2453184.5            180           1600              1.0
4             2453184.5            180           1600              1.0
...                 ...            ...            ...              ...
1880460       2457291.5            269           1726             13.0
1880461       2457300.5            278           0126              9.0
1880462       2457144.5            122           2052             13.0
1880463       2457309.5            287           2309             13.0
1880464       2457095.5             73           2128              9.0

          STAT_CAUSE_DESCR  CONT_DATE  CONT_DOY CONT_TIME  FIRE_SIZE  \
0            Miscellaneous  2453403.5      33.0      1730       0.10
1                Lightning  2453137.5     133.0      1530       0.25
2           Debris Burning  2453156.5     152.0      2024       0.10
3                Lightning  2453189.5     185.0      1400       0.10
4                Lightning  2453189.5     185.0      1200       0.10
...                    ...        ...       ...       ...        ...
1880460  Missing/Undefined  2457291.5     269.0      1843       0.01
1880461      Miscellaneous        NaN       NaN      None       0.20
1880462  Missing/Undefined        NaN       NaN      None       0.10
1880463  Missing/Undefined        NaN       NaN      None       2.00
1880464      Miscellaneous        NaN       NaN      None       0.10

        FIRE_SIZE_CLASS   LATITUDE   LONGITUDE  OWNER_CODE       OWNER_DESCR  \
0                     A  40.036944 -121.005833         5.0              USFS
1                     A  38.933056 -120.404444         5.0              USFS
2                     A  38.984167 -120.735556        13.0  STATE OR PRIVATE
3                     A  38.559167 -119.913333         5.0              USFS
4                     A  38.559167 -119.933056         5.0              USFS
...                 ...        ...         ...         ...               ...
1880460               A  40.481637 -122.389375        13.0  STATE OR PRIVATE
1880461               A  37.617619 -120.938570        12.0    MUNICIPAL/LOCAL
1880462               A  37.617619 -120.938570        12.0    MUNICIPAL/LOCAL
1880463               B  37.672235 -120.898356        12.0    MUNICIPAL/LOCAL
1880464               A  34.263217 -116.830950        13.0  STATE OR PRIVATE
```

```
        STATE COUNTY FIPS_CODE  FIPS_NAME  \
0          CA     63       063     Plumas
1          CA     61       061     Placer
2          CA     17       017  El Dorado
3          CA      3       003     Alpine
4          CA      3       003     Alpine
...       ...    ...       ...        ...
1880460    CA   None      None       None
1880461    CA   None      None       None
1880462    CA   None      None       None
1880463    CA   None      None       None
1880464    CA   None      None       None

                                                   Shape
0        b'\x00\x01\xad\x10\x00\x00\xe8d\xc2\x92_@^\xc0...
1        b'\x00\x01\xad\x10\x00\x00T\xb6\xeej\xe2\x19^\...
2        b'\x00\x01\xad\x10\x00\x00\xd0\xa5\xa0W\x13/^\...
3        b'\x00\x01\xad\x10\x00\x00\x94\xac\xa3\rt\xfa]...
4        b'\x00\x01\xad\x10\x00\x00@\xe3\xaa.\xb7\xfb]\...
...                                                  ...
1880460  b'\x00\x01\xad\x10\x00\x00P\xb8\x1e\x85\xeb\x9...
1880461  b'\x00\x01\xad\x10\x00\x00\x00\x80\xbe\x88\x11...
1880462  b'\x00\x01\xad\x10\x00\x00\x00\x80\xbe\x88\x11...
1880463  b'\x00\x01\xad\x10\x00\x00x\xba_\xaa~9^\xc0\xb...
1880464  b'\x00\x01\xad\x10\x00\x00\x1c\xa7\xe8H.5]\xc0...

[1880465 rows x 39 columns]
```

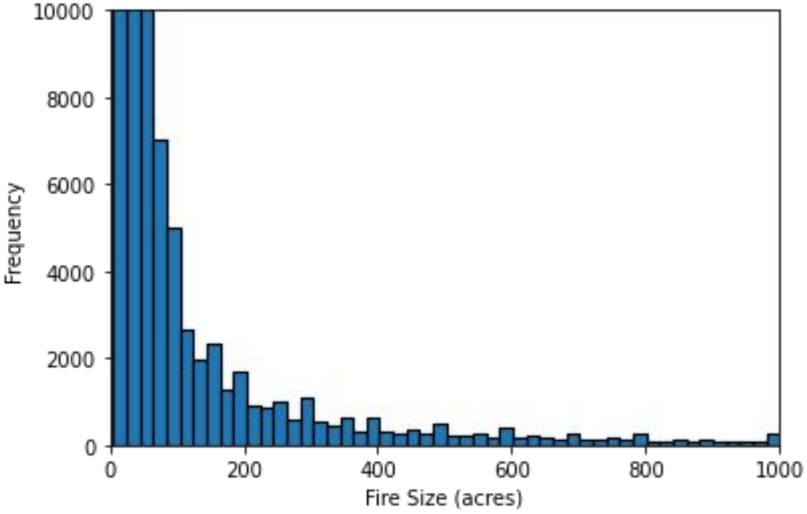
Initial goal (and how it failed)

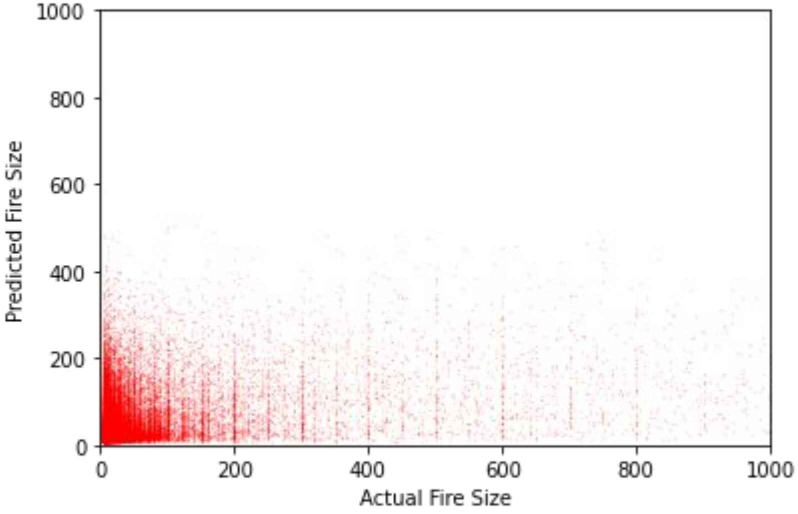
- Predict fire size from given latitude
- Predict fire size from relevant factors (lat, long, state, time of year)



- Used Knn
- Had average error of about half a standard deviation
- Source of error came from larger fires that were included

Initial goal (and how it failed)





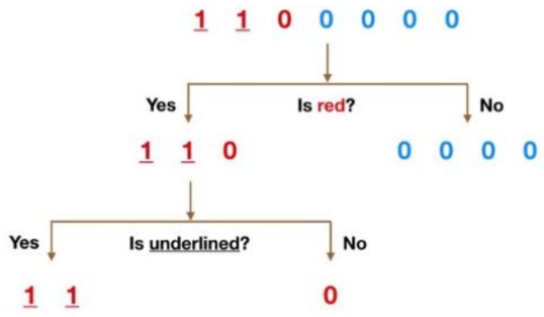
Main Issue: Heavy, Consistent Data skew to Smaller Fires

Random Forest Algorithm

- Decision Trees
  - At each node, the data is split into multiple, distinct groups.
  - Example: Decision tree with two features

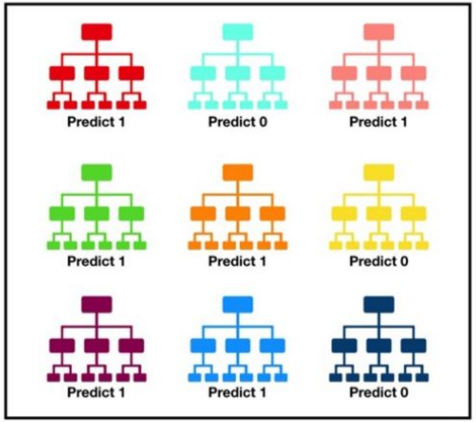
A Random Forest consists of a large number of decision trees that operate together

- The uncorrelated trees act as a committee
  - A few individual errors will not skew the prediction
  - Robust to outliers



Simple Decision Tree Example

Source: towardsdatascience.com

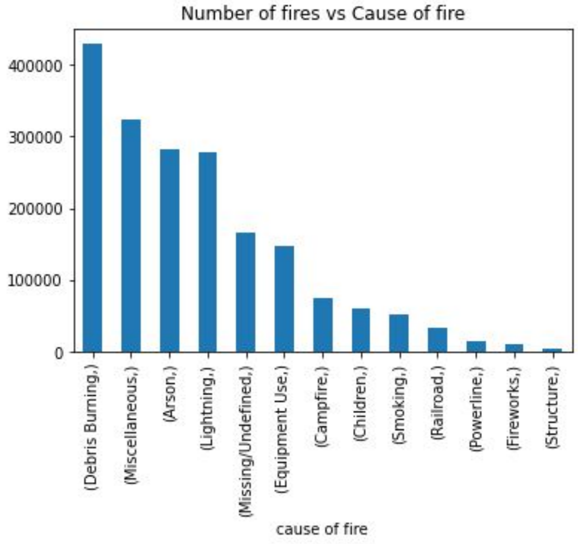


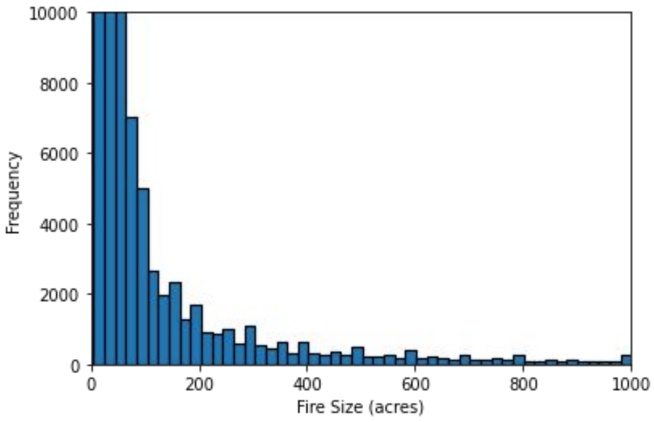
Tally: Six 1s and Three 0s
Prediction: 1

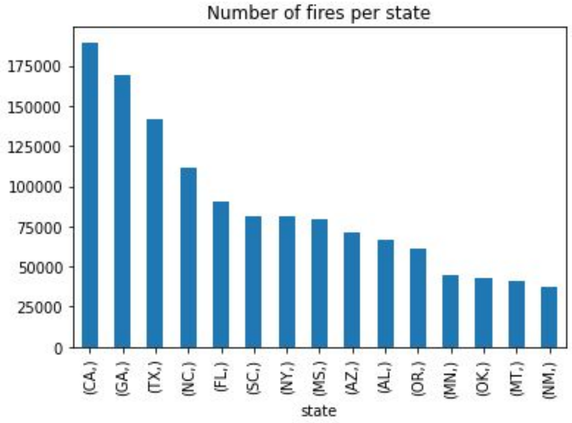
Source: towardsdatascience.com

Analyzing Data

- Using matplotlib to visualize our data
  - A significant amount of debris fires
  - Very few structure fires
  - Many fires are of unknown cause
  - An extremely large amount of smaller fires







Training Inputs

- Numerical data that described or provided information about the fire were kept.
- Columns that were empty were removed
- The labels were removed but saved to check our predictions

```

fire_info = pd.DataFrame(np.c_[df['FIRE_YEAR'],df['DISCOVERY_DATE'],df['DISCOVERY_DOY'],df['DISCOVERY_TIME'], df['STAT_CAUSE_CODE'], df['STAT_CAUSE_DESCR'],df['CONT_DATE'], \
                               df['CONT_DOY'],df['CONT_TIME'],df['FIRE_SIZE'],df['LATITUDE'],df['LONGITUDE'],df['OWNER_CODE']], \
                         columns= ['year', 'date', 'day of year', 'time', 'cause code', 'cause descr', 'containment date', 'containment doy', 'containment time', 'size', 'latitude', \
                                   'longitude', 'owner code'])

fire_info = fire_info.dropna()

fire_causes = pd.DataFrame(np.c_[fire_info['cause code'], fire_info['cause descr']], columns=['cause code', 'cause descr'])
fire_info = fire_info.drop(['cause code', 'cause descr'], axis=1)
```

First Results

ML Model:

- RandomForestClassifier
- 100 estimators (trees)
  - Doubling the estimators would only yielded a 0.2% increase in mean accuracy

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=100)
rf = rf.fit(X_train, y_train)
print(rf.score(X_test,y_test))
pred = rf.predict(X_test)
```

```
0.6284010268943173
```

Interpreting Results:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

target_names = ['Lightning', 'Equipment Use', 'Smoking', 'Campfire', \
                'Debris Burning', 'Railroad', 'Arson', 'Children', \
                'Miscellaneous', 'Fireworks', 'Powerline', 'Structure', \
                'Missing/Undefined']
conf_mat = confusion_matrix(y_test, pred)
print(classification_report(y_test, pred, target_names=target_names))
print(conf_mat)
```

Interpreting Results

- Classification Report
  - Recall for some causes are very high, some are very low
  - Some fire attributes might be too similar
- Confusion Matrix
  - Columns = predicted, rows = actual
  - Matches the classification report
  - Diagonals are the True Positives per feature

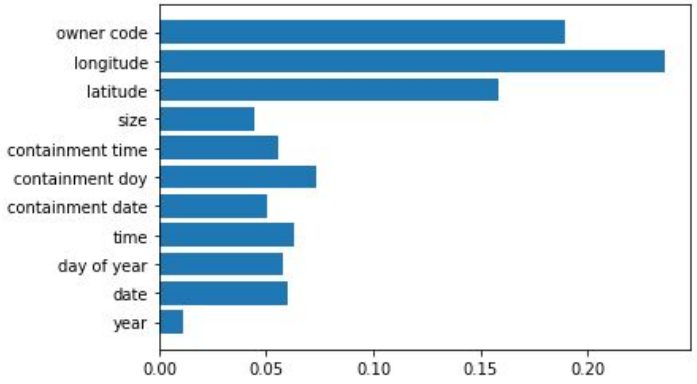
```
                   precision    recall  f1-score   support

         Lightning       0.77      0.91      0.84     43817
     Equipment Use       0.40      0.23      0.29      9955
           Smoking       0.31      0.06      0.10      4503
          Campfire       0.56      0.40      0.47     10417
    Debris Burning       0.56      0.70      0.62     34199
          Railroad       0.77      0.25      0.37      1522
             Arson       0.61      0.60      0.61     27906
          Children       0.44      0.22      0.29      5379
     Miscellaneous       0.54      0.56      0.55     29671
         Fireworks       0.64      0.53      0.58      2040
         Powerline       0.38      0.07      0.12      1596
         Structure       0.38      0.07      0.11       505
 Missing/Undefined       0.80      0.79      0.80      6892

          accuracy                           0.63    178402
         macro avg       0.55      0.41      0.44    178402
      weighted avg       0.61      0.63      0.61    178402

[[39940   177    22   560   836     6   546    51  1478    57    13     1
    130]
 [ 1419  2246    76   311  2332    12   935   140  2199    44    38    10
    193]
 [  651   213   260   354  1215     5   433    77  1201    26     4     1
     63]
 [ 2737   217    60  4190  1340     9   560    53  1184    27    10     1
     29]
 [ 1103   591   105   469 23899    37  4460   443  2730   104    37    14
    207]
 [  246    55    17    41   464   376   158     9   132     5     2     0
     17]
 [ 1028   506    76   313  5630    14 16782   298  3001   101    25     1
    131]
 [  398   174    43   121  1792     3   673  1164   852    84     6    14
     55]
 [ 3152  1067   130  1114  4278    23  2377   267 16575   151    48     8
    481]
 [  260    47    13    17   187     5   184    45   182  1083     1     3
     13]
 [  268   142    10    19   434     0   173    23   382     5   118     1
     21]
 [   35    30     1     8   184     0    65    40    94     6     3    33
      6]
 [  433    98    14    27   228     1   116     6   512     9     6     0
   5442]]
```

Improving Results

- Removing feature of least importance
- Grouping up similar labels

```
print(rf.feature_importances_)

fire_features = fire_info.columns.to_list()
plt.barh(fire_features, rf.feature_importances_)

```

```
[0.0111788  0.06029448 0.0579831  0.06271325 0.05018791 0.0730023
 0.05591178 0.04463047 0.15837646 0.2361124  0.18960905]
<BarContainer object of 11 artists>
```




```
unknown = ['Miscellaneous', 'Missing/Undefined'] # 1
unintentional = ['Railroad', 'Powerline', 'Structure'] # 2
preventable = ['Debris Burning', 'Children', 'Equipment Use', 'Campfire', 'Smoking', 'Fireworks'] # 3
natural = ['Lightning'] # 4
purposeful = ['Arson'] # 5
```

Testing Set Results

- The mean accuracy increased from 62.8% to 70.9%.
- f1-scores are consistently higher in all categories except ‘Unintentional’.

```
1 rf = ske.RandomForestClassifier(n_estimators=100)
2 rf = rf.fit(X_train, y_train)
3 print(rf.score(X_test,y_test))
4 pred = rf.predict(X_test)
```

```
0.7087981076445331
```

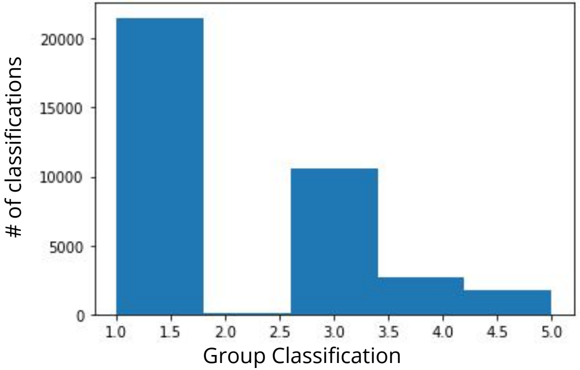
```
               precision    recall  f1-score   support

      Unknown       0.66      0.59      0.62     36563
Unintentional       0.72      0.12      0.21      3623
  Preventable       0.67      0.76      0.71     66493
      Natural       0.82      0.89      0.85     43817
   Purposeful       0.68      0.54      0.60     27906

     accuracy                           0.71    178402
    macro avg       0.71      0.58      0.60    178402
 weighted avg       0.71      0.71      0.70    178402

[[21438    42 10556  2756  1771]
 [  483   439  2073   413   215]
 [ 6223    95 50705  4680  4790]
 [ 1514    14  3046 38862   381]
 [ 2747    19  9373   760 15007]]
```

Validation Set Results

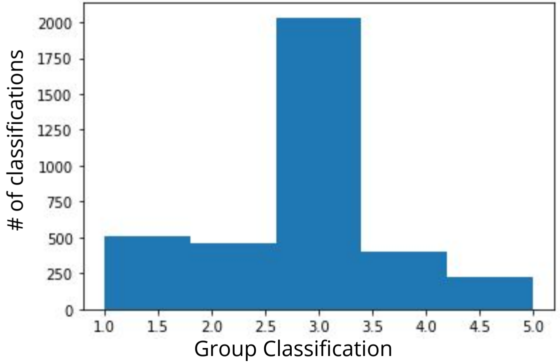
- Our validation set results are extremely similar to our testing set results, which is a good thing.
- Most of the labels that were incorrectly classified were classified to group 3, or fires under the 'Preventable' group (shown in the histograms).

	precision	recall	f1-score	support
Unknown	0.66	0.59	0.62	36617
Unintentional	0.74	0.12	0.20	3662
Preventable	0.67	0.76	0.71	66425
Natural	0.82	0.88	0.85	43757
Purposeful	0.67	0.53	0.60	27941
accuracy			0.71	178402
macro avg	0.71	0.58	0.60	178402
weighted avg	0.71	0.71	0.70	178402

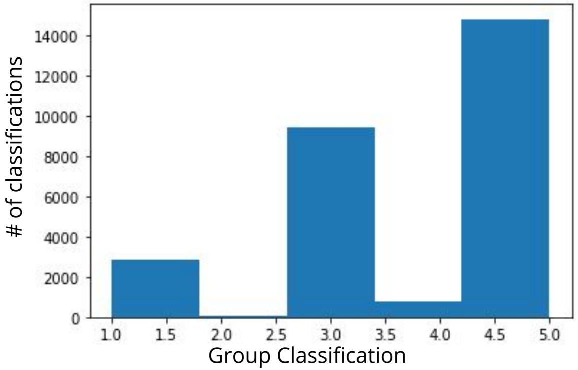
21651	44	10356	2790	1776
466	435	2113	395	253
6202	73	50519	4776	4855
1604	13	3100	38683	357
2819	21	9346	818	14937



Group 1



Group 3



Group 5

Analysis of Results

- Our model was able to predict the cause of the fire based on fire attributes, with a decent accuracy.
- Where our model struggles categorizing: ‘Preventable’ vs ‘Unintentional’
  - Varying fire characteristics
- Improving results: predict for smaller regions

What was Learned?

- Organizing your data can be challenging, especially if it is skewed
- Using different models and approaches within Machine Learning can greatly affect your results
- It is important to understand your data before trying to apply a model