**Analysis of Human-Caused Forest Fires**

* **By Surbhi Anand**

**TABLE OF CONTENTS**

LIST OF FIGURES……………………………………………………………………………4

LIST OF TABLES…………………………………………………………………………….5

1. INTRODUCTION………...………………………………………………………………...7

1.1 *Problem Background* ……………………………………………………………...8

1.2 *Motivation* ………………………………………………………………………...8

1.3 Problem Statement ……………………………………………………………......9

2. REVIEW OF LITERATURE...............................................................................................10

3. TERMINOLOGIES.............................................................................................................11

4. SOLUTION METHODOLOGY..........................................................................................12

4.1 *Objective* ................................................................................................................12

4.2 *Data Preparation & Processin*g ...........................................................................13

4.3 *Exploratory Data Analysis* ....................................................................................15

4.3.1 Structure of data......................................................................................15

4.3.2 Line Chart................................................................................................16

4.3.3 Bar Chart.................................................................................................17

4.3.4 Heat Map.................................................................................................21

4.3.5 Maps........................................................................................................21

4.3.6 Clustering................................................................................................23

4.4 *Evaluation Metrics of Classification Problems*.....................................................27

4.4.1 Accuracy..................................................................................................28

4.4.2 Precision..................................................................................................28

4.4.3 Recall.......................................................................................................29

4.4.4 Specificity................................................................................................29

4.4.5 False Positive Rate (FPR).......................................................................29

4.4.6 False Negative Rate (FNR).....................................................................29

4.4.7 ROC.........................................................................................................29

4.4.8 AUROC/ AUC........................................................................................29

4.5 *Machine Learning Methods*...................................................................................30

4.5.1 Naïve Bayes.............................................................................................30

4.5.2 Multinomial Logistic Regression............................................................30

4.5.4 Ensemble Methods..................................................................................30

4.5.4.1 *Bagging*....................................................................................30

4.5.4.2 *Random Forests*........................................................................31

4.5.4.3 *Boosting*....................................................................................31

4.5.3 KNN........................................................................................................32

5. ANALYSIS & RESULTS....................................................................................................34

5.1 Data Partition for Machine Learning......................................................................34

5.2 Multi-class ROC.....................................................................................................34

5.3 Application of Machine Learning Models.............................................................35

5.3.1 Model 1: Naïve Bayes.............................................................................37

5.3.2 Model 2: Multinomial Logistic Regression........ ...................................37

5.3.3 Model 3: Random Forests.......................................................................37

5.3.4 Model 4: Boosting...................................................................................40

5.3.5 Model 5: KNN.........................................................................................42

6. CONCLUSIONS..................................................................................................................44

7. RECOMMENDATIONS FOR FUTURE ...........................................................................46

8. REFERENCES.....................................................................................................................47

9. APPENDIX..........................................................................................................................50

9.1 Appendix A: Output...............................................................................................50

9.2 Appendix B: R Code..............................................................................................52

9.3 Appendix C: Links to other files............................................................................80

**LIST OF FIGURES**

Figure 1: Structure of dataset 16

Figure 2: US Forest fires during 2010-18 16

Figure 3: US human-caused forest fires during 2010-18 17

Figure 4: Seasonal count of US human forest fires during 2010-18 18

Figure 5: Monthly count of forest fires during 2010-18 18

Figure 6: Count of US human forest fires w.r.t cause & season during 2010-18 19

Figure 7: Count of US human forest fires on weekday/weekend during 2010-18 19

Figure 8: Count of US human forest fires on weekday/weekend w.r.t cause during 2010-18 20

Figure 9: Count of human forest fires w.r.t fire class during 2010-18 20

Figure 10: Count of US human forest fires w.r.t class and cause during 2010-18 21

Figure 11: US forest fires count map during 2010-18 22

Figure 12: US B Class human forest fires count map 22

Figure 13: US A Class human forest fires count map 23

Figure 14: Clustering of US states for 2018 25

Figure 15: Count of US forest fires state wise for 2019 26

Figure 16: Bagging 31

Figure 17: Boosting 32

Figure 18: Correlation between quantitative attributes 35

Figure 19: OOB plot for random forest 38

Figure 20: plot of OOB vs mtry 39

Figure 21: Variable Importance Plot for Random Forest Model 40

Figure 20: Selection of k 42

**LIST OF TABLES**

Table 1: Columns in Final Dataset 14

Table 2: US States grouped in clusters based on forest fire risk for 2018 25

Table 3: Confusion Matrix 28

Table 4: Top 3 affected states by each cause 35

Table 5: Partitions in dataset 35

Table 6: Columns used in model creation 36

Table 7: Parameter of Random Forest model 39

Table 8: Parameter of Boosting model 41

Table 9: AUROC of models 43

**ABSTRACT**

Forest fires have become a major concern worldwide. Frequent occurrences have affected the environment. Millions of acres burning every year have led to substantial economic and human losses. Lightning and volcanic eruptions are causes of natural forest fires, which account for 10% of the total number of fires (Wildfires. III., 2021, U.S. Department of the Interior, 2018). The remaining 90% of forest fires are the result of sheer human negligence. Common causes of human-caused forest fires are smoking, campfire, debris burning, and equipment (Grala, K., 2017).

Over the years, machine learning techniques have improved fire management through prediction models. The impact of biophysical factors on natural forest fires is a well-researched area (B. K. Singh, et al., 2019, D. Rosadi et al., 2020). But human forest fires were not studied in the context of socio-economic factors, such as poverty and population until recently (Grala, K., 2017, Rodrigues, M., 2014, Sturtevant, B. R., 2007).

This project analyzed human-caused forest fires in the United States from 2010-2018. The analysis process included two components: Exploratory data analysis and machine learning algorithms: Naïve Bayes, Multinomial Logistic Regression, Random Forests, Boosting and KNN.

Results suggested that socioeconomic factors affect the occurrence of different forest fires with different human ignition causes. This analysis can offer valuable insights in the context of forest fire predictions and management.

**INTRODUCTION**

***Problem Background***

Forest fires are the most frequent disaster all over the world. Although they occur worldwide, they are a growing problem in the United States. In 2010, around 11M acres were burned in California. Forest fires not only damage the environment but also result in substantial economic and human losses. In 2020, the annual suppression cost for US wildfires was more than 2 billion (Borrelli, L., 2021). More than 2 million properties are at risk in California only. Around 33,000 people die every year globally because of forest fires (Reinberg, S, 2021). With climate change, the problem has become worse.

It is imperative to develop an understanding of forest fire behavior patterns. There are two categories of forest fires, natural and human. Natural wildfires are caused by lightning or volcanic eruptions. Common causes of human-caused forest fires are smoking, campfire, debris burning, and equipment (Grala, K., 2017). In the US, around 80% to 90% of fires are attributed to human activities.

***Motivation***

Many organizations actively collect and analyze data to strategize fire management. United States Department of Agriculture (USDA) issues “Wildfire Hazard Potential” to categorize regions based on risk. It highlights the areas at high risk of forest fires. National Interagency Coordination Center (NICC) also issues “Significant Wildfire potential outlook” every month to improve preparedness against the devastating consequences of forest fires. Machine learning (ML) algorithms are applied on collected data to develop forest fire predictive models (Xie, Y., 2018). Many journals have discussed the application of ML techniques, such as random forests, ensemble methods, SVM, and regression in this field. However, the application of ML in the context of different causes of human-caused forest fires is a recent development.

Studies (Grala, K., 2017, Rodrigues, M., 2014, Sturtevant, B. R., 2007) have now started analyzing the impact of socio-economic factors such as population, the unemployment rate on human forest fire occurrences of different origins. The motivation of this project is to make some contribution towards the improvement of human-caused wildfire management strategies. Results can suggest how prevention strategies could be customized to a specific human ignition cause of the forest fire.

***Problem Statement***

This project addresses the problem of identifying if there is an impact of human factors on the occurrences of human forest fires. Data analysis was used to approach the problem. The study was exclusively done on human-caused forest fires during 2010-18. The forest fire dataset was cross-referenced with socio-economic data (Short, K. C, 2021, USDA ERS). Exploratory data analysis revealed some useful insights. Machine learning algorithms were applied to the dataset to predict the cause of human forest fires.

**REVIEW OF LITERATURE**

Machine learning techniques have been applied in wildfire science and management since the early 1990s. The emergence of powerful computing ability enabled the application of ML in this domain. ML techniques are used to predict forest fire occurrence, susceptibility, risk and impact.

Earlier research (B. K. Singh, 2019, D. Rosadi, 2020) focussed on how biophysical factors, such as climate, fuel, and soil moisture, influence forest fires. Ensemble methods (Xie, Y, 2018), regression, and SVM predicted burned area with reasonable accuracy. While this prior work improved the understanding of forest fires, the impact of socio-economic factors was neglected.

The inclusion of human factors in forest fire prediction is now being explored (Calef, M. P.,2008, Narayanaraj, G., 2012). Some studies broadly categorized wildfires as human-caused and natural (Stephens, 2005). A few of them only focussed on arson wildfires in the context of socio-economic factors (Prestemon and Butry, 2005, Prestemon and Donovan, 2008). Different human ignition causes such as smoking or recreational activity are grouped into one category in some research work (Prestemon et al., 2010).

A recent study suggests that the fire mitigation approach should vary as per the type of human-caused fire (Martinez, 2009). Another study (Syphard, A.D., 2015) suggests that the number of fires and areas burned vary with different ignition causes. A small amount of literature (Grala, K., 2017, Rodrigues, M., 2014, Sturtevant, B. R., 2007) examines how socio-economic factors affect the occurrence of different forest fires with different human ignition causes in one or a few states in the United States.

**TERMINOLOGIES**

**AUROC/ AUC:** Area under the curve (AUC) of ROC gives a single value which measures the performance of a prediction model.

**IDE**: Integrated Development Environment is a software to build and test code. For example: PyCharm, RStudio.

**ML**: Machine Learning algorithms learn from data and build models to make predictions.

**NICC**: National Interagency Coordination Center allocates resources to manage wildfires across the United States.

**OOB:** Out-of-bag (OOB) error is the average error in classifying or predicting samples that are not present in the respective training sample of the tree.

**ROC:** Receiver Operating Characteristics (ROC) curve is curve plotted between true positive rate and false positive rate to measure the performance of a predictive model.

**SVM:** Support Vector Machine is a supervised machine learning algorithm to perform regression and classification.

**USDA:** United States Department of Agriculture is a federal department founded in 1862, responsible for making laws for agriculture, forestry, and food quality.

**WUI**: Wildlife Urban Interface is the meeting point of human developed areas and undeveloped wildland.

**SOLUTION METHODOLOGY**

Data Analysis has emerged as a powerful technique, applicable in all domains. The key idea is to derive meaningful insights from data which ultimately leads to informed decision-making. The process involves the following phases (Asati, A., 2021):

* Identifying objectives or goals
* Collecting and storing data from multiple sources
* Cleaning, organizing and transforming raw data
* Analysing data using analytics tools (Excel, SQL, R and Tableau)
* Sharing insights with others

This project also followed the same life cycle of data analysis process. Following Data Analytics tools were used for the project:

Microsoft Excel: It is a spreadsheet software program for analysing and visualizing data.

SQLite DB Browser: It is an open-source tool to read and manipulate data files through queries.

R Studio: It is an open-source tool for statistical analysis. It provides an integrated development environment (IDE) for programming in R.

Tableau: It is one of the leading visualization tools, helping people to explore data and share insights from it using different kinds of charts, plots, and graphs.

Programming Languages: SQL, R

***Objective***

The objective of the project was to conduct data analysis on human-caused forest fires. This project aimed to:

* Explore prior research work on Forest Fire prediction.
* Correlate forest fire data with socio-economic factors.
* Create forest fire prediction models and assess their performance.
* Perform clustering of forest fire incidents in the context of socio-economic factors for the United States.
* Identify states with high risk of forest fires
* Derive meaningful insights from the analysis and suggest measures for fire management.

***Data Preparation & Processing***

Data were collected from U.S. government websites and archives (Short, K. C, 2021, USDA ERS). The USDA Research Data Archive has maintained a database of wildfires that occurred in the U.S from 1992 to 2018. Forest fires data was obtained from the USDA archives. 2010 U.S. The data was initially read using SQLite DB Browser. Using SQL queries, data was filtered for 2010-18 period and then exported to CSV format.

The inspiration for collecting socio-economic data was obtained from prior work in this area (Grala, K., 2017). Census was the pivotal data source for socio-economic data. Population, poverty count & rate, unemployment count & rate, housing units, and median household income were collected at the county level from 2010 to 2018. All the datasets were cleaned, filtered, and transformed using Microsoft Excel.

Following steps were taken for data preparation:

* Filtered forest fire records for the 2010-2018 period.
* Filtered forest fire records of human origins.
* Dropped descriptive columns of forest fire dataset
* Purged forest fire records with missing causes of human-caused forest fires.
* Purged forest fire records with unknown locations.
* Filtered socio-economic data for the 2010-2018 period.

The compiled dataset was created by the union of forest fire and socio-economic datasets. The final dataset comprised 21 columns and 443,129 rows. Among 21 attributes, 11 of them were quantitative, and the rest were either descriptive or categorical. Following were the columns in compiled dataset:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| STATE | Character | Name of the state (in abbreviated form) |
| COUNTY | Character | Name of the county |
| LATTUDE | Numeric | Latitude of fire origin location |
| LONGITUDE | Numeric | Longitude of fire origin location |
| FIRE\_YEAR | Numeric | Year of forest fire occurrence |
| DISCOVERY\_DATE | Date | Date of forest fire occurrence |
| DISCOVERY\_MONTH | Character | Month of forest fire occurrence |
| DISCOVERY\_DAY | Character | Day of the week |
| WEEKEND\_FLAG | Boolean | Boolean flag for weekend |
| SEASON | Character | Spring/ Summer/ Fall/ Winter |
| NWCG\_CAUSE\_  CLASSIFICATION | Character | Natural/ Human |
| NWCG\_GENERAL\_CAUSE | Character | Different causes of human-caused forest fires |
| FIRE\_SIZE | Numeric | Size of the fire (in hectares) |
| FIRE\_SIZE\_CLASS | Character | Size of wildfire:  Class A - one-fourth acre or less;  Class B - more than one-fourth acre, but less than 10 acres;  Class C - 10 acres or more, but less than 100 acres;  Class D - 100 acres or more, but less than 300 acres;  Class E - 300 acres or more, but less than 1,000 acres;  Class F - 1,000 acres or more, but less than 5,000 acres;  Class G - 5,000 acres or more. |
| HOUSING\_UNITS | Numeric | Number of housing units |
| POPULATION | Numeric | Population estimates |
| POVERTY\_RATE | Numeric | Poverty rate |
| POVERTY\_COUNT | Numeric | Number of people living under poverty |
| UNEMPLOYMENT\_COUNT | Numeric | Number of unemployed |
| UNEMPLOYMENT\_RATE | Numeric | Unemployment rate |
| MEDIAN\_INCOME | Numeric | Median household income |

*Table 1: Columns in Final Dataset*

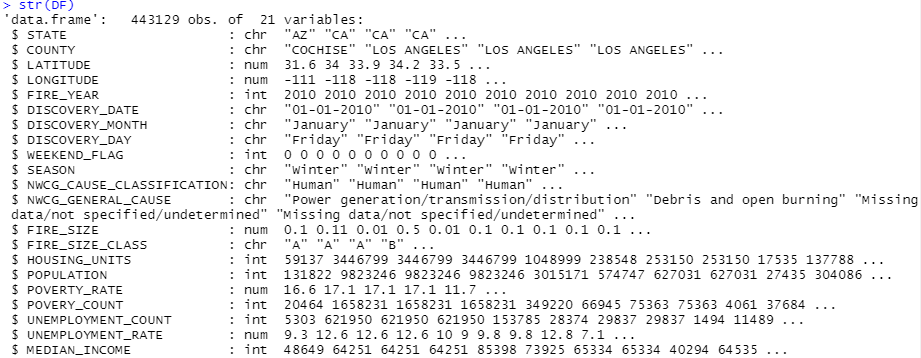
***Exploratory Data Analysis***

Exploratory Data Analysis (EDA) is used to make sense of the data collected. It is an approach to identify patterns or anomalies at the outset of the project. Visualizations, like bar chart, line chart are often used to perform EDA.

**Structure of data.** Understanding the structure of dataset is the first step towards data analysis.

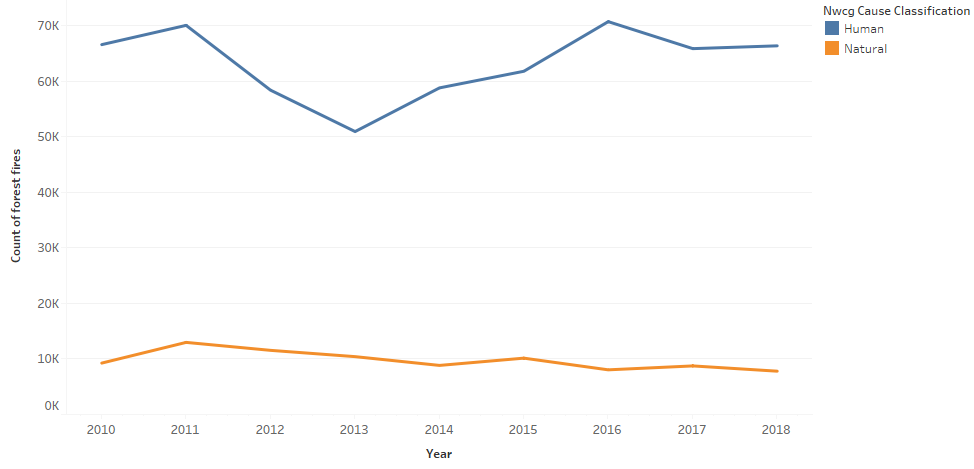
* The dataset had 443,129 observations of 21 attributes. One of them was response variable and the rest were predictive attributes.
* The response variable, i.e., NWCG\_GENERAL\_CAUSE was categorical, with 10 classes: Power generation/transmission/distribution, Debris and open burning, Misuse of fire by a minor, Smoking, Arson/incendiarism, Equipment and vehicle use, Recreation and ceremony, Fireworks, Railroad operations and maintenance, and Firearms and explosives use
* None of the columns had any missing or null values.

Figure 1 showed the structure of the dataset. The number of records and attributes, column names, and their data type were a part of the output.



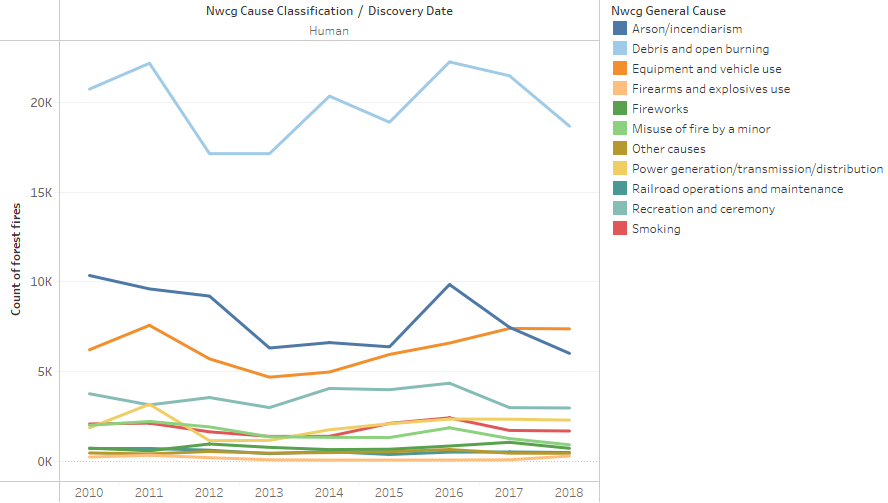
*Figure 1: Structure of dataset*

**Line Chart.** A line chart represents change in quantitative values over a period of time. Following line charts were created using Tableau.



*Figure 2: US Forest fires during 2010-18*

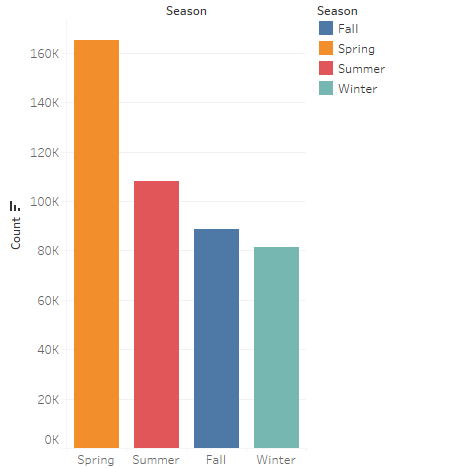
Figure 2 showed that the number of natural forest fires had gradually decreased over the years. However, the number of human-caused forest fires was considerably higher than that of natural forest fires during 2010-18. Since 2013, the number of human-caused forest fires has followed an increasing trend.



*Figure 3: US human-caused forest fires during 2010-18*

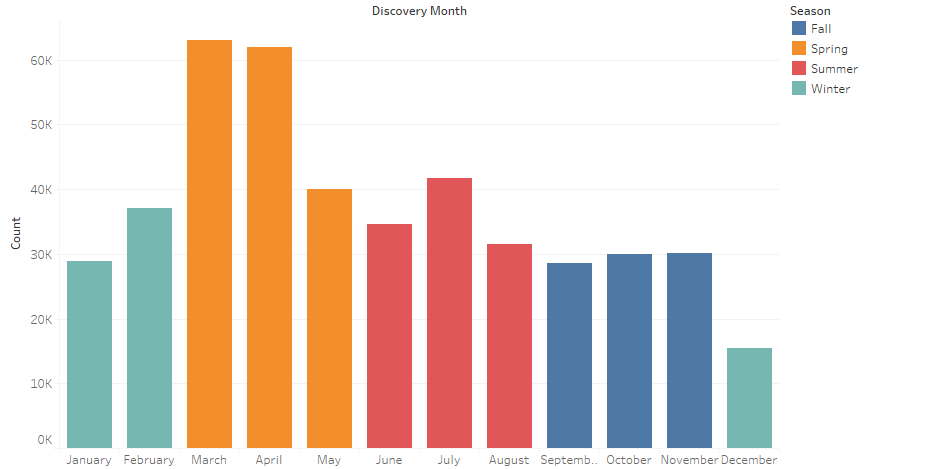
The line chart, Figure 3 showed that among all the causes of human forest fires, “Debris and open burning” had maximum instances, followed by “Arson/ incendiarism” and “Equipment and Vehicle use.” The count of all the types of human forest fires fluctuated heavily over the 2010-18 time period.

**Bar Graph.** A bar graph plots categorical data with rectangular bars. The bars could be horizontal or vertical, with length or height proportional to the value they represent. Following bar charts were created by Tableau.



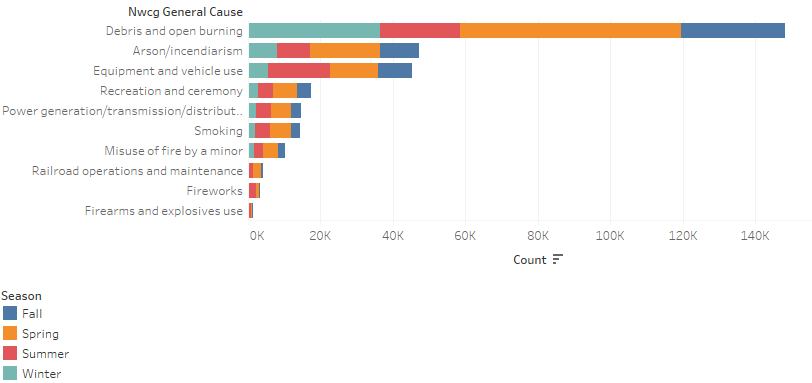
*Figure 4: Seasonal count of US human forest fires during 2010-18*

Figure 4 showed that the maximum number of human-caused forest fires occurred in Spring Season, followed by Summer Season during 2010-18.



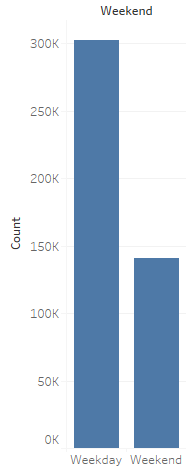
*Figure 5: Monthly count of forest fires during 2010-18*

Figure 5 showed that during 2010-18, March, April (spring season), and July (summer season) had the maximum occurrences of human-caused forest fires.

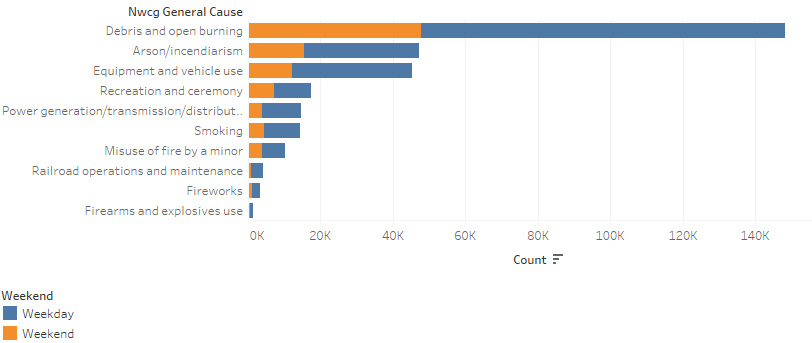


*Figure 6: Count of US human forest fires w.r.t cause & season during 2010-18*

The bar plot in Figure 6 showed that” Debris and open burning” had caused maximum number of forest fires in 2010-18. The plot also showed that all the causes had higher instances in Spring season.

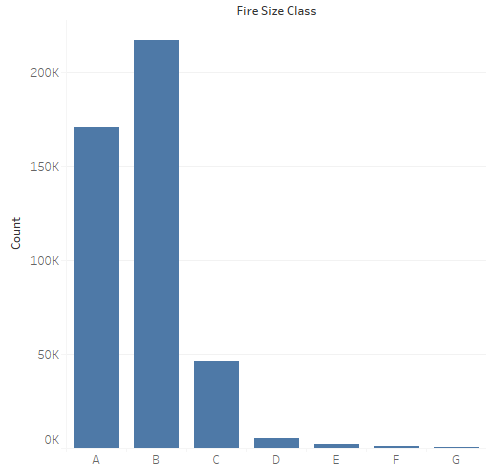


*Figure 7: Count of US human forest fires on weekday/weekend during 2010-18*



*Figure 8: Count of US human forest fires on weekday/weekend w.r.t cause during 2010-18*

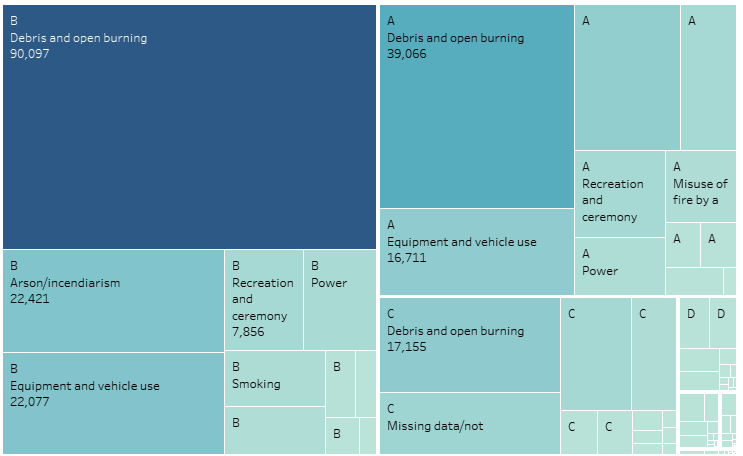
The above bar plots Figure 7 & Figure 8 showed that human forest fires occurred more on weekday. For all the causes, weekday had more instances of fires as compared to weekend.

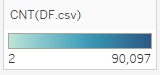


*Figure 9: Count of human forest fires w.r.t fire class during 2010-18*

Figure 9 showed that maximum number of human-caused forest fires were of B class (wildfire size of 0.25 to 10 acre) and A class (wildfire size of 0.25 acre or less).

**Heatmap.** In the case of a heat map, numeric values are translated to color. It visualizes any metrics or phenomenon as color in two dimensions. The variation in color shows the reader the spots or clusters where the intensity of the metrics is high. Following heat map was created in Tableau.

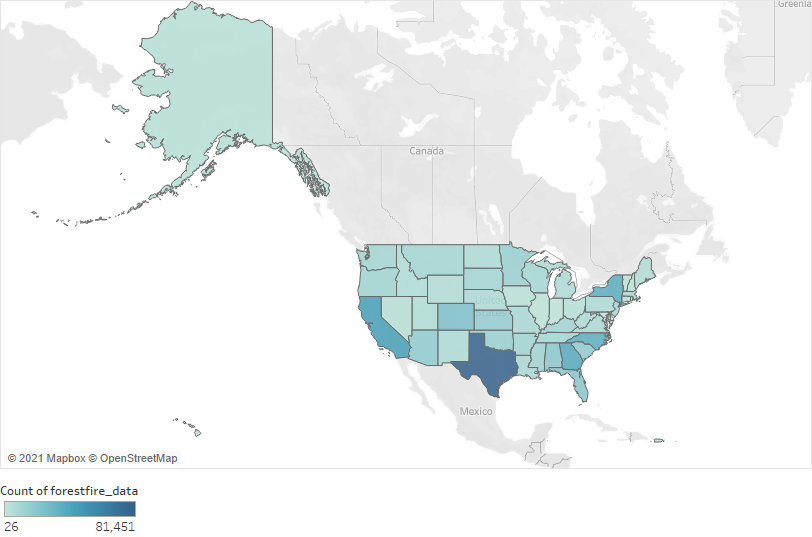




*Figure 10: Count of US human forest fires w.r.t class and cause during 2010-18*

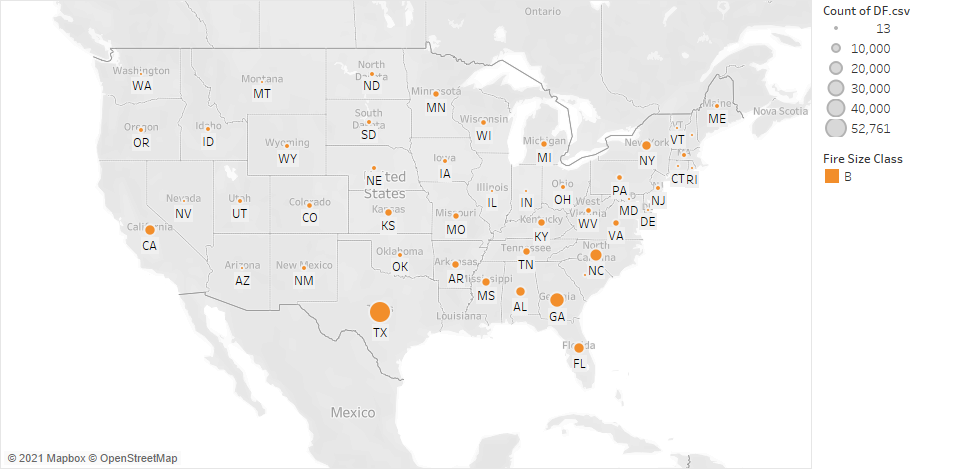
The heatmap, Figure 10 showed that among A class and B class forest fires, “Debris and open burning” was the major cause, followed by “Equipment and vehicle use”.

**Maps.** Maps visually represent geographical data. It adds an aesthetic element to data representation. Also, maps display the distribution of data in each region through the color intensity, shape, or size of data points.

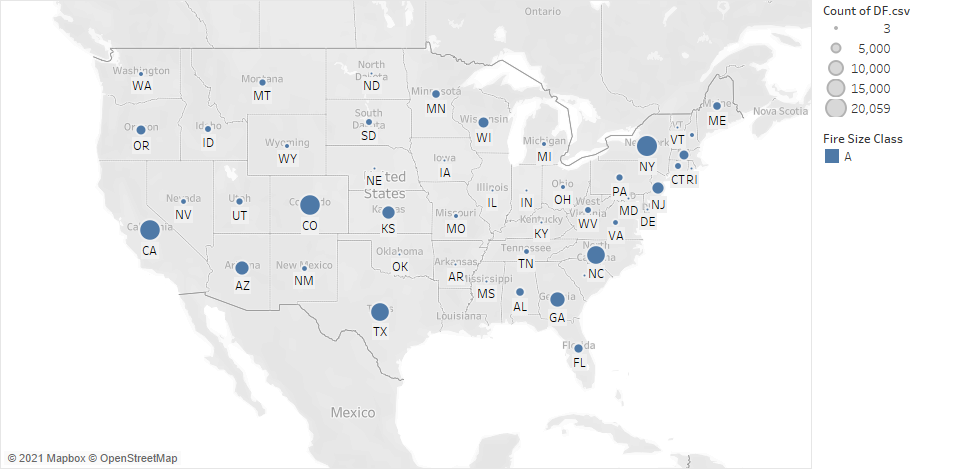


*Figure 11: US forest fires count map during 2010-18*

Figure 11 plotted all the occurrences (natural and human) of forest fires over 2010-18. The intensity of color reflected the number of fire instances in a state. Texas, California, Georgia, New York, and North Carolina have had the most wildfires.



*Figure 12: US B Class human forest fires count map*



*Figure 13: US A Class human forest fires count map*

The above map plots, Figure 12 and Figure 13, showed the distribution of human forest fires across US states during 2010-18. The size of the data point reflected the count of forest fires in a state. The maximum number of B-class forest fires (wildfire size of 0.25 to 10 acre) occurred in Texas, California, and Georgia. While, A class forest fire (wildfire size of 0.25 acre or less) occurred in Texas, California, Colorado, New York, North Carolina, and Arizona.

**Clustering of fire-prone states.** Clustering puts data points into clusters based on similarity. From a dataset perspective, two similar records are grouped in a cluster, while dissimilar records are put in different clusters. It is an unsupervised machine learning method because the model is not trained beforehand. Customer segmentation is one of the applications of clustering.There are two types of clustering:

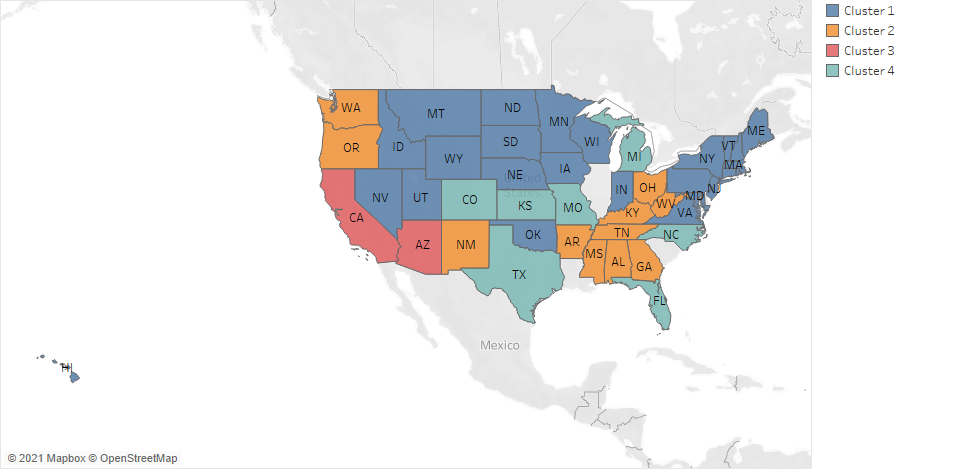
***Hierarchical clustering*.** It starts with n (number of data points) clusters and sequentially merge similar clusters until one cluster is left based on the distance between clusters. It is computationally intensive, which is why it is not suitable for large datasets. Following are the steps in Hierarchical clustering:

* Start with n (number of data points) clusters.
* Compute distance matrix.
* Merge closest clusters and update distance matrix.
* Repeat the process until one cluster is left
* Dendrogram captures the sequential merge process.
* Computationally intensive, not suitable for large datasets.

***K-means clustering*.** It requires pre-specified number of clusters. It groups data points in such a way that minimizes dispersion with cluster. It is a simple algorithm and less computationally intensive. Following are the steps in K means clustering:

* Start with a pre-specified number of clusters.
* Compute centroid of each cluster.
* Compute the distance of each data point from each centroid.
* Re-assign data points to the closest cluster.
* Iterate until no reassignment of data points is required.
* Simple algorithm, computationally less intensive.

Tableau uses K means clustering to perform clustering. It automatically identifies the optimum number of clusters to group data.

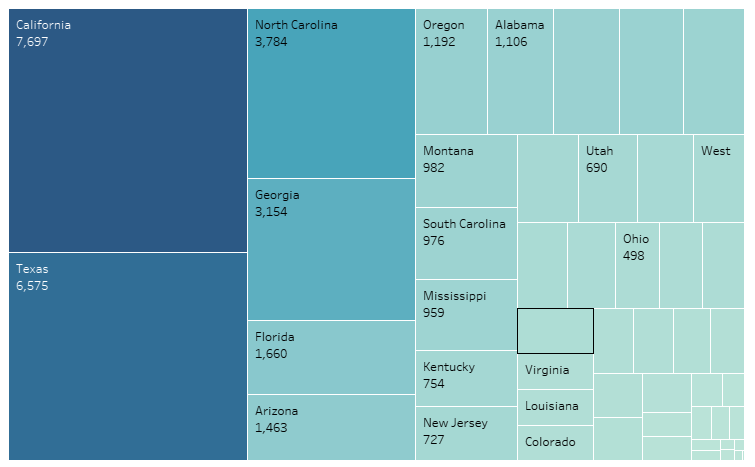


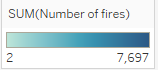
*Figure 14: Clustering of US states for 2018*

Figure 14 showed that Tableau automatically grouped all the states of the US in 4 clusters based on forest fires and socio-economic data for 2018. States, categorized in clusters, are summarized in Table 2.

|  |  |
| --- | --- |
| **Cluster** | **States** |
| Cluster 4 (high risk) | Texas, Colorado, Kansas, Michigan, North Carolina, Montana |
| Cluster 3 | California, Arizona |
| Cluster 2 | Washington, Oregon, New Mexico, Arkansas, Mississippi, Alabama, Georgia, Tennessee, Kentucky, Ohio, West Virginia |
| Cluster 1 (low risk) | Remaining states |

*Table 2: US States grouped in clusters based on forest fire risk for 2018*





*Figure 15: Count of US forest fires state wise for 2019*

In 2019, California, Texas, North Carolina, Georgia, Florida, and Arizona were the most affected by human-caused wildfires (Borrelli, L., 2021). Clustering, based on 2018 statistics, also suggested a similar outcome. Texas, Florida, North Carolina were grouped in cluster 1, while California and Arizona were grouped in cluster 2. The states in clusters 1 and 2 had seen the maximum instances of human forest fires in 2018. Alignment in the high-risk states for 2018 and 2019 reinforced the correlation between forest fires and socio-economic factors.

***Evaluation Metrics for Classification models***

Once the data is prepared, it is fed into a classification model. Suppose there are two classes, 0 and 1, with the class of interest as 1. A classification model assigns a class label to input data. The model generates output in terms of the probability of belonging to the class of interest. If the probability is more than 0.5, the outcome would be class 1, else 0. A confusion matrix draws a picture of how well the model is performing. Table 3 shows the confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Actual Class** | |
|  |  | Class 1 | Class 0 |
| **Predicted Class** | Class 1 | True Positive | False Positive |
| Class 0 | False Negative | True Negative |

*Table 3: Confusion Matrix*

The definitions of terminologies used in the confusion matrix are as follows:

True Positive: When the model predicted class 1, and it is true.

False Positive: When the model predicted class 1, but it is false.

False Negative: When the model predicted class 0, but it is false.

True Negative: When the model predicted class 0, and it is true.

Many classification metrics are derived from the confusion matrix. The most common one is accuracy.

**Accuracy.** Accuracy is the percentage of cases classified correctly.

(1)

But in case, the dataset is imbalanced, having a dominant class, accuracy is not a suitable metric to assess model performance. The model can simply classify all the records in the data set as the dominant class, and yet achieve a good accuracy score. For imbalanced data set, instead of accuracy, other metrics, such as Area under Receiver Operating Characteristic (AUROC), are used (Brownlee, J., 2021).

Finer- grained measures that can be derived from confusion matrix are as follows:

**Precision.** Out of all positive predictions by the model, how many were actually positive class.

(2)

**Recall.** Out of all positive classes, how many were predicted correctly by the model. It is also called as True Positive Rate (TPR) and Sensitivity.

(3)

**Specificity.** Out of all negative classes, how many were predicted correctly by the model. It is also called True Negative Rate (TNR).

(4)

**False Positive Rate (FPR).** Out of all negative classes, how many were predicted incorrectly by the model. It is equal to (1- Specificity).

(5)

**False Negative Rate (FNR).** Out of all positive classes, how many were predicted incorrectly by the model. It is equal to (1- Sensitivity).

(6)

**ROC.** The Receiver Operating Characteristics (ROC) curve is plotted between TPR and FPR for different threshold values.

**AUROC/ AUC.** The area under the curve (AUC) of ROC quantifies the model performance. The higher the AUC, the better is the model performance. An AUC of more than 0.5 suggests that the model is working. An AUC of 0.5 indicates that model is not good at distinguishing classes. An AUC of less than 0.5 means that model is performing worse than making random guesses. It is also called as AUROC (Area under Receive Operating Characteristics Curve).

In the project dataset, "Debris an open burning" accounted for 48% of the total number of forest fires, creating an imbalance. So, the AUROC was used to compare the performance of different machine learning models.

***Machine Learning Methods***

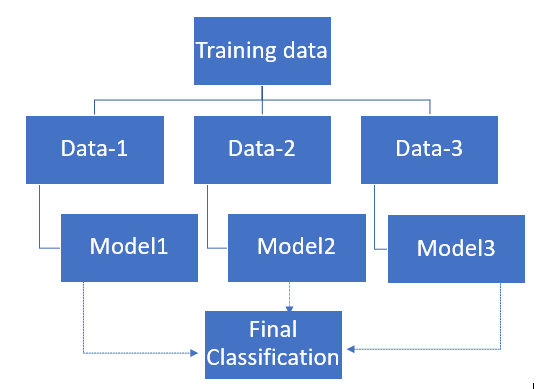
Five machine learning algorithms were used to classify causes of human forest fires based on socio-economic factors.

**Naïve Bayes.** It is a simple and fast classification technique, based on Bayes’ Theorem. It uses conditional probability of observing a class given a set of features. This algorithm assumes statistical independence of probability of occurrence of each feature. This model can work as a baseline model.

**Multinomial Logistic Regression.** It is used for multi-class classification problem. It internally works similar to logistic regression. It uses log odd of the outcomes to determine the probability of class. For k classes, the algorithm creates k-1 models. For example, when there are three classes, A, B and C, the models created are: A vs (B, C) and B vs (A, C).

**Ensemble Methods.** They combine the results from multiple machine learning models to improve the predictive accuracy. The key idea is that insights drawn from multiple models are more likely to be accurate than a single model. They have been considered as the most important development in the field of data mining. They include the following algorithms.

***Bagging*.** Bagging term is coined by combining words bootstrap and aggregating. In bootstrapping, smaller datasets are created from training data set through sampling with replacement. Models are created for each smaller dataset. Now, the test data set is applied on all the models. Outcome classes of each model are voted, and final classification is created based on the highest voted outcome class.Using bootstrapping to create uncorrelated models, and then aggregating their results is called bagging. Figure 16 shows the process involved in bagging.

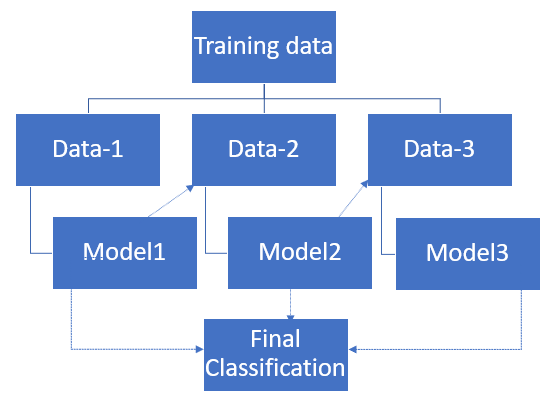


*Figure 16: Bagging Method (created in word using Smart Art)*

***Random Forests*.** Random forest is an extension of bagging where each model is a decision tree. Also, feature randomness is introduced by limiting the number of variables a tree can consider for splitting. Since the decision trees are created on a smaller dataset with a limited number of features, it takes less time to train the models. So, creating a large number of trees can is not computationally intensive.

***Boosting*.** It differs from random forest in the way each model is created and how the results of individual models are combined. It boosts the models sequentially so that each new model is an improvement on the previous model.

Gradient boosting is one of the Boosting techniques. This technique uses decision tree as model. Initially, all the observations in the data set are given equal weightage and first base model is created on the data set. The observations which are misclassified by the base model are then given more weightage as compared to the observations which are correctly classified. The second decision tree is run on the weighted data. The process runs for pre-specified number of iterations. For the test data prediction, weighted sum of individual classifiers is the outcome. Figure 17 shows the process involved in boosting.



*Figure 17: Boosting Method (created in word using Smart Art)*

**KNN.** It is based on a very simple algorithm. It classifies or predicts a value for a new data point, based on the k nearest points. In case of prediction, the average outcome of nearest neighbours is the end result. For classification, the majority class among k nearest neighors is the final outcome.

Nearest neighbor is identified using distance metric. One common method of computing distance between data points is Euclidean distance.

D = √ (x1 – y1) 2 + ...... (xn – yn) 2  (5)

Categorical attributes need to be handled carefully as distance cannot be computed for them. So, they are converted into numeric equivalents.

Identifying optimal k is also important for the model. Less value of k can lead to overfitting, while large value can smoothen the noise completely. As the number of k increases, accuracy improves initially. But, after a certain threshold, starts dropping. The best practice is to compute model accuracy for a range of values of k, and select the optimal value.

**ANALYSIS & RESULTS**

The analysis of forest fire data identified three most affected states for each cause.

|  |  |
| --- | --- |
| **Human Forest Fire Causes** | **3 Most Affected States** |
| Arson/ incendiarism | New York, North Carolina, Georgia |
| Debris and Open Burning | Texas, Georgia, North Carolina |
| Equipment and Vehicle Use | Texas, California, North Carolina |
| Firearms and explosive use | Idaho, California, Utah |
| Fireworks | WY, WV, NC |
| Misuse of fire by minor | California, North Carolina, Georgia |
| Power generation/ transmission/ distribution | Texas, California, North Carolina |
| Railroad operations and maintenance | Texas, Georgia, Wisconsin |
| Recreation and ceremony | Georgia, Kentucky, Mississippi |
| Smoking | New York, Texas, California |

*Table 4: Top 3 affected states by each cause*

***Data Partition for Machine Learning***

Data set was partitioned into three sets – training, validation and test data set. Below is the percentage split for each partition.

|  |  |
| --- | --- |
| **Partition** | **Percentage** |
| Training | 60% |
| Validation | 20% |
| Test | 20% |

Table 5: Partitions of dataset

Training data set was used for training the machine learning models. Validation dataset was used for validating the results and tuning the model parameters to get better results. Test data was used to test the model and get final results.

***Multi-class ROC***

ROC curve is extended from binary classification to multiclass ROC. The idea is to take pair wise classes and find the area under the curve (AUC) for each pair. The paper (Hand, D.J, 2001) proposes overall AUC as average of AUC of each pair of classes.

(6)

In Equation 6,

L = number of classes

ci = class i

cj = All the remaining classes

= Area under curve of ROC of ci and cj.

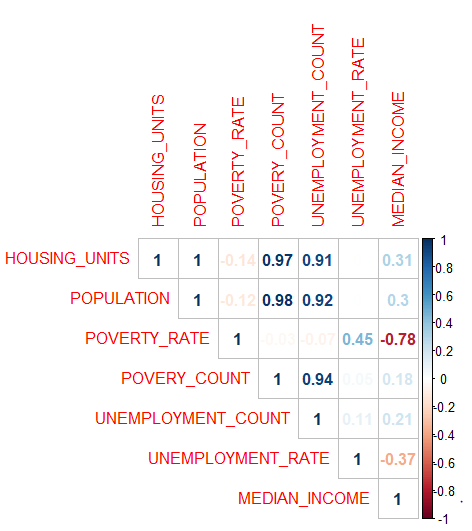
If there are L classes, then number of pairs are L(L-1)/2 classes.

***Application of Machine Learning Models***

All the models were coded in R programming language. The code is included in Appendix B, with separate sections for each model.

The compiled data set was imported in R. Descriptive attributes were dropped as they cannot be used for model creation. Columns dropped were: STATE, COUNTY, FIRE\_YEAR, DISCOVERY\_DATE, NWCG\_CAUSE\_CLASSIFICATION.

Correlation among quantitative fields was calculated. Figure 18 showed the correlation among quantitative fields. The function “findCorrelation” was used to identify highly correlated columns. They were “HOUSING\_UNITS,” “POPULATION,” and “POVERTY\_COUNT.” The best practice to deal with correlated columns is to drop and check if they impact the model performance. If model performance improves after dropping highly correlated columns, they should be dropped (BUS 235C). Otherwise, they can be included in the model creation. For each ML algorithm, two models were created, with and without highly correlated columns.



*Figure 18: Correlation between quantitative attributes*

Following columns were included in model creation for each algorithm.

|  |  |
| --- | --- |
| **Independent Variables** | LATITUDE, LONGITUDE, DISCOVERY\_MONTH, DISCOVERY\_DAY, WEEKEND\_FLAG, SEASON, FIRE\_SIZE, FIRE\_SIZE\_CLASS, POVERTY\_RATE, UNEMPLOYMENT\_COUNT, UNEMPLOYMENT\_RATE,  MEDIAN\_INCOME  HOUSING\_UNITS, POPULATION, POVERTY\_COUNT  (Highly correlated columns) |
| **Dependent Variable** | NWCG\_GENERAL\_CAUSE |

*Table 6: Columns used in model creation*

NWCG\_GENERAL\_CAUSE represents ten different human origins of human forest fires. Based on forest fire and socio-economic data, models are supposed to classify fires of different origins. It is a multi-class classification problem. AUROC is an important evaluation metric to compare the performance of models for such problems (Narkhede, S., 2021). AUROC was computed for all the models to compare the performance among models.

**Model 1: Naïve Bayes**

Naïve Bayes is a simple and fast algorithm. It does not have any parameters that can be tuned to improve the predictive power of the model. When all the attributes were included in model creation, AUC was 0.6303. When highly correlated columns were dropped, AUC was 0.6517. The output is shown in Figure A-1. Both the models yielded an AUROC of higher than 0.5, which means the model can classify different origins of human forest fires based on the data. Naïve Bayes algorithm does not have any method to compute significance of attributes (Reddy, S. K., 2020)

**Model 2: Multinomial Logistic Regression**

2-tail z test was used to identify significant columns in the dataset. All the predictor attributes were significant. When all the attributes were included in model creation, AUC was 0.6439. When highly correlated columns were dropped, AUC was 0.6375. The results were in agreement with 2-tail z test. Dropping columns reduced the AUC value. Both the models yielded an AUROC of higher than 0.5, which means the model can classify different origins of human forest fires based on the data.

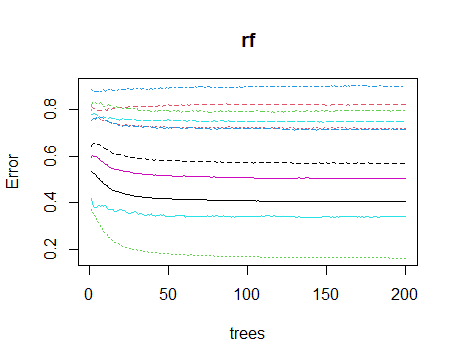
**Model 3: Random Forests**

Out-of-bag (OOB) error is the average error in classifying or predicting samples that are not present in the respective training sample of the tree. Ntree (Number of trees) was determined by checking the OOB error rate of the random forest model.

First a model was created with an initial value of ntree as 500. Internally, R module trains different random forest models with increasing number of trees. For each random forest model, OOB error is calculated. By plotting OOB error against the count of trees, an optimal value of ntree is determined.

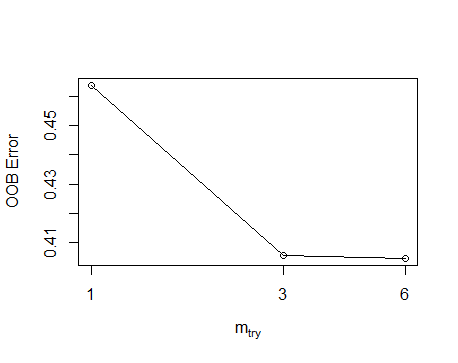
Each curve represents OOB error curve for a class of response variable. In the project dataset, NWCG\_GENERAL\_CAUSE is the response variable, with ten type of human forest fire causes. So, ten curves are generated.

Figure 19 shows that the error rate for each class of response variable became stable near 200. So, the value for ntree was chosen as 200.



*Figure 19: OOB plot for Random Forest*

 In random forest, each decision tree is trained and tested with a limited number of predictor variables. Number of predictor variables used at each decision tree is mtry. First multiple random forest models are trained for different values of mtry. The resultant OOB error for each random forest model is calculated and plotted against corresponding mtry value. Mtry was determined by tuneRF function in R. The function selects mtry such that OOB error is minimal. Figure 20 shows that the minimum OOB error was achieved at mtry =6.

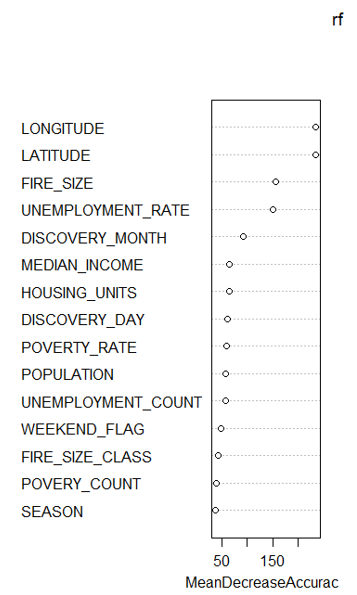


*Fig 20: plot of OOB error vs mtry*

|  |  |  |
| --- | --- | --- |
| **Parameter Name** | **Parameter Description** | **Value** |
| ntree | Number of trees | 200 |
| mtry | Number of variables randomly sampled as candidates at each split | 8 |
| importance | This parameter indicates if importance of predictors be assessed | True |

Table 6: Parameter of Random Forest model

When all the attributes were included in model creation, AUC was 0.8293. When highly correlated columns were dropped, AUC was 0.8289. The output is shown in Figure A-3. Both the models yielded an AUROC of higher than 0.5. The AUC value suggests that there is approximately 82% chance that model will distinguish between different causes of human forest fires.



*Figure 21: Variable Importance Plot of Random Forest Model*

Figure 21 showed the importance of variables for Random Forest model. Most important attribute were: LONGITUDE, LATITUDE, FIRE\_SIZE, UNEMPLOYMENT\_RATE, DISCOVERY\_MONTH and MEDIAN\_INCOME.

**Model 4: Gradient Boosting (XG Boost)**

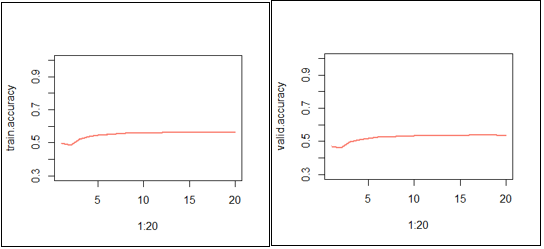
XGBoost model was tuned using the following parameters. Selection criteria determines how parameter values were selected.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter Name** | **Parameter Description** | **Range** | **Value** | **Selection Criteria** |
| nrounds | Boosting iterations |  | 500 |  |
| max\_depth | Max Tree Depth. Larger is the depth, more complex is the model which leads to higher chances of overfitting. |  | 3 | To keep the depth of the tree minimal, 3 was used. Lower depth reduces the probability of overfitting |
| eta | Shrinkage controls learning rate. Lower eta leads to slower computation. | 0 to 1 | 0.3 | Default value of eta used by XgBoost model in R is 0.3.  eta as 0.1 was also used, but this increased the runtime with no improvement in AUC. |
| gamma | Minimum Loss Reduction. This parameter prevents overfitting. Higher is the gamma, lower is the difference in train and test cross validation | 0 to infinity | 0.01 | Default value of gamma = 0 used by XGBoost model in R. As max\_depth is low, not much overfitting is expected. So, gamma was assigned a small value of 0.01 |
| colsample\_bytree | Subsample Ratio Columns. This parameter controls number of features supplied to a tree. | 0 to 1 | 1 | Default value of colsample\_bytree is 1. 1 means 100% of columns are supplied to a tree. |
| min\_child\_weight | Minimum Sum of Instance Weight | 0 to infinity | 1 | Default value of this parameter is 1. If the weight of a leaf node in decision tree is less than 1, then recursive partitioning is stopped. |
| subsample | Subsample Percentage supplied to a tree | 0 to 1 | 1 | Default value is 1. By keeping value as 1, each decision tree is trained with all the records of the data set, though predictor variables may vary for each model. |

*Table 7: Parameters for Boosting model*

When all the attributes were included in model creation, AUC was 0.8298. When highly correlated columns were dropped, AUC was 0.8275. The output is shown in Figure A-4. Both the models yielded an AUC of higher than 0.5, indicating that the models have good classification power. There is approximately 82% of chance that Boosting model will classify correctly the different causes of human forest fires. The important attributes were Latitude, Longitude, Fire\_Size, Housing\_units, Unemployment\_count.

**Model 5: KNN**



*Figure 20: Selection of k value*

Figure 20 shows the process to select k (number of neighbors). The accuracy was tested for k ranging from 0 to 20. The training set yielded maximum accuracy, i.e., 56% for k = 16. For validation set, maximum accuracy, i.e., 53% was obtained for k = 20. So, the final model was created with 20 neighbors. The model made predictions for testing data set with accuracy of 53%.

When all the attributes were included in model creation, AUC was 0.7113. When highly correlated columns were dropped, AUC was 0.7107. The output is shown in Figure A-5. Both the models yielded an AUC of higher than 0.5. There is approximately 71% of chance that KNN model will classify correctly the different causes of human forest fires.

AUROC was used for evaluating the performance of different models. Performance of different machine learning models on testing data set is summarized in Table 8.

|  |  |  |
| --- | --- | --- |
| **Machine Learning Models** | **Area under curve**  **(With highly correlated columns)** | **Area under curve**  **(Without highly correlated columns)** |
| Naïve Bayes | 0.6303 | 0.6517 |
| Multinomial Logistic Regression | 0.6439 | 0.6375 |
| Random Forests | 0.8293 | 0.8289 |
| Gradient Boosting | 0.8298 | 0.8275 |
| KNN | 0.7113 | 0.7107 |

*Table 8: AUC of models*

The performance of both the ensemble methods was better than the rest of the algorithms. AUROC for both the ensemble methods was approximately 82%. It means that there is an 82% chance that models correctly classify the different causes of human forest fires. Between the two ensemble methods, Random Forests is recommended ML algorithm as it is not computationally intensive. Boosting algorithm takes eight to twelve hours for execution, while Random Forests run within five minutes. KNN also performed well, with an AUC of 0.71. But the execution time was about twelve hours. Also, for four out of five algorithms, a drop was observed when highly correlated columns were not included in model creation. So, they should be taken into consideration when model is created.

**CONCLUSIONS**

Over the past few years, there have been many instances of human-caused fires. If their occurrences can be modelled and predicted, the forest department can prepare well ahead accordingly.

Data analysis has proven its utility in this domain. Reliable prediction models have improved the efficiency of forest management. This project analyzed different causes of human forest fires in the context of poverty, population, housing, median income, and unemployment. The project can be summarized as below.

* Exploratory data analysis offers valuable insights. As opposed to common belief, chances of forest fires are more on a weekday, rather than a weekend. Similarly, human forest fires are more likely to occur in the spring and summer seasons. Additional resources should be allocated to the fire department in these seasons by the government.
* Debris and open burning pose the highest threats. An effective plan to clear out debris in the fall season can reduce the number of incidents.
* Understanding the different origins of human forest fires is critical. Mitigation strategies can be customized as per the plausible cause. Also, awareness programs can be launched to educate people regarding the most probable causes of forest fires in their county.
* Clustering identifies states that are at high risk of forest fires. A high-end surveillance system, including satellites, sensors, and scanners can help monitor and respond immediately in case of any incidents. Texas, Colorado, Kansas, Michigan, North Carolina, Montana, California, and Arizona were found to be the risk-prone states.
* A high value of AUC for Ensemble methods highlights the impact of socio-economic factors on different human forest fires.
* The results of EDA and ML algorithms are aligned. EDA showed how number of fires varied across different seasons. Season also appeared as an important factor for Random Forest model. Similarly, fire size was an important feature for Random Forest model as well as EDA. All the map plots showed location is an important factor for forest fires. Clustering highlighted risk prone states. On the same lines, Latitude and longitude were the most important factors for Random Forest model.
* Both components of data analysis, EDA and ML algorithms suggest that forest fires are affected by temporal (day, month and season), spatial (latitude and longitude) and socio-economic (population, unemployment, poverty and median income) factors.

**RECOMMENDATIONS FOR FUTURE**

The lack of standard datasets is a challenge for applying machine learning algorithms in the wildfire. The scope of available public datasets is limited to a small area, for example, a national park or some state. A well-structured and current dataset of forest fires can help improve the analysis.

The inclusion of the following spatial attributes can improve understanding of human-caused forest fires.

* Distance between forest fire origin location to the nearest road
* Distance between forest fire origin location to the nearest railway track
* Distance between forest fire origin location to the nearest city

Wildland urban interface (WUI) are fire prone areas (Idriz, B., 2021). Dataset can be more informative if WUI areas can be identified for each state. Due to the unavailability of data, they were not included in the analysis of this project.

A comprehensive dataset, combining both meteorological and socio-economic data of forest fires can produce a panoramic picture of this hazard.

Deep Learning techniques can be used to improve the predictive power of models.

**REFERENCES**

Alabama median household income (in 2018 dollars), 2014-2018 by County. (n.d.). Retrieved November 15, 2021, from <https://www.indexmundi.com/facts/united-states/quick-facts/alabama/median-household-income#table>.

Asati, A. (2021, March 1). *Data Analysis Process: 8 useful phases of data analysis process*. EDUCBA. Retrieved November 26, 2021, from https://www.educba.com/data-analysis-process/.

Borrelli, L. (n.d.). *U.S. wildfire statistics*. Bankrate. Retrieved November 26, 2021, from https://www.bankrate.com/insurance/homeowners-insurance/wildfire-statistics/.

Bureau, U. S. C. (2021, October 26). *National, state, and County Housing Unit Totals: 2010-2019*. Census.gov. Retrieved November 15, 2021, from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-total-housing-units.html.

Brownlee, J. (2021, April 30). *Tour of evaluation metrics for imbalanced classification*. Machine Learning Mastery. Retrieved November 27, 2021, from https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.

Calef, M. P., McGuire, A. D., & Chapin, F. S. (2008). Human influences on wildfire in Alaska from 1988 through 2005: An analysis of the spatial patterns of human impacts. *Earth Interactions*, *12*(1), 1–17. <https://doi.org/10.1175/2007ei220.1>

*Download Data*. USDA ERS - Download Data. (n.d.). Retrieved September 29, 2021, from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

*Facts + statistics: Wildfires*. III. (2021, April 28). Retrieved October 7, 2021, from https://www.iii.org/fact-statistic/facts-statistics-wildfires.

Grala, K., Grala, R. K., Hussain, A., Cooke, W. H., & Varner, J. M. (2017). Impact of human factors on wildfire occurrence in Mississippi, United States. *Forest Policy and Economics*, *81*, 38–47. <https://doi.org/10.1016/j.forpol.2017.04.011>

Hand, D.J., Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning* 45**,**171–186 (2001). <https://doi.org/10.1023/A:1010920819831>

*Human-caused wildfires*. National Interagency Fire Center. (n.d.). Retrieved October 7, 2021, from https://www.nifc.gov/fire-information/statistics/human-caused.

Idriz, B., Wang, J., & Brown-Stevens, A. (2021, April 27). *WUI: Three letters at the core of Wildfire, housing, and land use*. Greenbelt Alliance. Retrieved November 29, 2021, from https://www.greenbelt.org/blog/wui-three-letters-at-the-core-of-wildfire-housing-and-land-use/.

Martinez, J., Vega-Garcia, C., Chuvieco, E., 2009. Human-caused wildfire risk rating for prevention planning in Spain. J. Environ. Manag. 90, 1241–1252.

*Multiclass.roc: Multi-class auc*. RDocumentation. (n.d.). Retrieved November 20, 2021, from <https://www.rdocumentation.org/packages/pROC/versions/1.18.0/topics/multiclass.roc>.

Narkhede, S. (2021, June 15). *Understanding AUC - roc curve*. Medium. Retrieved November 27, 2021, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

Narayanaraj, G., & Wimberly, M. C. (2012). Influences of forest roads on the spatial patterns of human- and lightning-caused wildfire ignitions. *Applied Geography*, *32*(2), 878–888. <https://doi.org/10.1016/j.apgeog.2011.09.004>

Prestemon, J.P., Butry, D.T., 2005. Time to burn: modeling wildland arson as an autoregressive crime function. Am. J. Agric. Econ. 87, 756–770.

Pham, B. T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H. P., Phong, T. V., Nguyen, D. H., Le, H. V., Mafi-Gholami, D., Prakash, I., Thi Thuy, H., & Tuyen, T. T. (2020). Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry*, *12*(6), 1022. <https://doi.org/10.3390/sym12061022>

Prestemon, J.P., Donovan, G.H., 2008. Forecasting resource-allocation decisions under climate uncertainty: fire suppression with assessment of net benefits of research. Am. J. Agric. Econ. 90, 1118–1129.

Prestemon, J.P., Butry, D.T., Abt, K.L., Sutphen, R., 2010. Net benefits of wildfire prevention education efforts. For. Sci. 56, 181–192.

Reinberg, S. (n.d.). *Wildfires cause more than 33,000 deaths globally each year ...* USNews. Retrieved November 26, 2021, from https://www.usnews.com/news/health-news/articles/2021-09-09/wildfires-cause-more-than-33-000-deaths-globally-each-year.

Rodrigues, M., & de la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, *57*, 192–201. <https://doi.org/10.1016/j.envsoft.2014.03.003>

Reddy, S. K. (2020, October 31). *Feature importance in naive Bayes classifiers*. InBlog. Retrieved December 2, 2021, from https://inblog.in/Feature-Importance-in-Naive-Bayes-Classifiers-5qob5d5sFW.

Rosadi, D., Andriyani, W., Arisanty, D. and Agustina, D., "Prediction of Forest Fire Occurrence in Peatlands using Machine Learning Approaches," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 48-51, doi: 10.1109/ISRITI51436.2020.9315359.

Singh, B.K., Kumar, N. and Tiwari, P., "Extreme Learning Machine Approach for Prediction of Forest Fires using Topographical and Metrological Data of Vietnam," *2019 Women Institute of Technology Conference on Electrical and Computer Engineering (WITCON ECE)*, 2019, pp. 104-112, doi: 10.1109/WITCONECE48374.2019.9092926

Short, K. C. (n.d.). *Spatial wildfire occurrence data for the United States, 1992-2018 [fpa\_fod\_20210617] (5th edition)*. Home. Retrieved September 29, 2021, from <https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5>.

Stephens, S.L., 2005. Forest fire causes and extent on United States Forest Service lands. Int. J. Wildland Fire 14, 213–222.

Sturtevant, B. R., & Cleland, D. T. (2007). Human and biophysical factors influencing modern fire disturbance in northern Wisconsin. *International Journal of Wildland Fire*, *16*(4), 398. <https://doi.org/10.1071/wf06023>

Syphard, A.D., Keeley, J.E., 2015. Location, timing and extent of wildfire vary by cause of ignition. Int. J. Wildland Fire 24, 37–47.

Xie, Y., & Peng, M. (2018). Forest fire forecasting using ensemble learning approaches. *Neural Computing and Applications*, *31*(9), 4541–4550. <https://doi.org/10.1007/s00521-018-3515-0>

U.S. Department of the Interior. (n.d.). *Wildfire causes and evaluations (U.S. National Park Service)*. National Parks Service. Retrieved October 7, 2021, from https://www.nps.gov/articles/wildfire-causes-and-evaluation.htm.

**Appendix**

**Appendix A: Output**

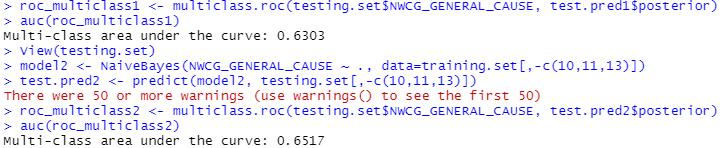


Figure A-1: Naïve Bayes Output

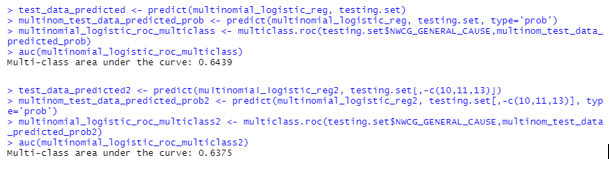


Figure A-2: Multinomial Logistic Regression Output

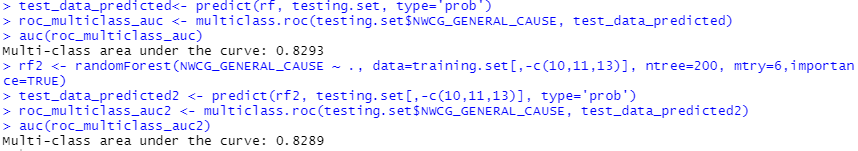


Figure A-3: Random Forests Output

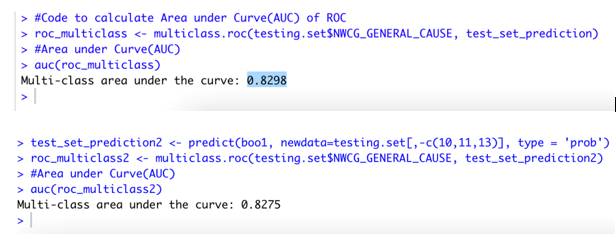


Figure A-4: Boosting Output

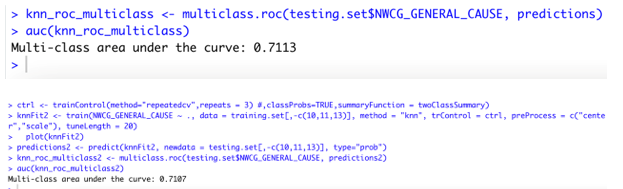


Figure A-5: KNN Output

**Appendix B: R Code**

***Pre-processing of Data***

*# Reading all the data files*

*# Filtering forest fire data*

*# Joining forest fire data with socio-economic data*

library(dplyr)

library(DBI)

library(RSQLite)

library(sos)

**if** (!require("pacman")) install.packages("pacman","sos")

pacman::p\_load(pacman,rio,tidyverse,sos)

mydb <- dbConnect(RSQLite::SQLite(), "FPA\_FOD\_20210617.sqlite")

table\_info <- dbGetQuery(mydb, 'PRAGMA table\_info(fires)')

table\_info

result <- dbGetQuery(mydb,'select

FIRE\_YEAR,

DISCOVERY\_DATE,

DISCOVERY\_DOY,

DISCOVERY\_TIME,

NWCG\_CAUSE\_CLASSIFICATION,

NWCG\_GENERAL\_CAUSE,

FIRE\_SIZE,

FIRE\_SIZE\_CLASS,

LATITUDE,

LONGITUDE,

STATE,

COUNTY

from fires

where

fire\_year>=2010 and fire\_year<=2018

and NWCG\_CAUSE\_CLASSIFICATION="Human"

and COUNTY != "NA"')

head(result,n=7)

dbDisconnect(mydb)

fileName <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level", "forestfire\_data.csv")

ret <- write.csv(x=result, fileName)

print(ret)

*## read data for Population estimates*

df <- import("Population\_2010-18.xls")

head(df,5)

num\_rows <- nrow(result)

county\_population <- vector(, num\_rows)

nrow(county\_population)

forest\_fire\_data <- data.frame(

fire\_data <- result,

population = county\_population

)

*#data clean up step*

cleanup\_forest\_fire\_data <- subset(forest\_fire\_data, !grepl("[0-9]+",forest\_fire\_data$COUNTY))

cleanup\_forest\_fire\_data$COUNTY <- toupper(cleanup\_forest\_fire\_data$COUNTY)

cleanup\_forest\_fire\_data$COUNTY2 <- gsub( "COUNTY","",cleanup\_forest\_fire\_data$COUNTY)

cleanup\_forest\_fire\_data$COUNTY <- cleanup\_forest\_fire\_data$COUNTY2

fileName <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level", "clean\_forestfire\_data.csv")

ret <- write.csv(x=cleanup\_forest\_fire\_data, fileName)

print(ret)

head(cleanup\_forest\_fire\_data,5)

data\_len <- nrow(cleanup\_forest\_fire\_data)

print(data\_len[1])

population\_data\_by\_year\_2010 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2010,

FIRE\_YEAR = 2010

)

head(population\_data\_by\_year\_2010,5)

population\_data\_by\_year\_2011 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2011,

FIRE\_YEAR = 2011

)

head(population\_data\_by\_year\_2012,5)

population\_data\_by\_year\_2012 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2012,

FIRE\_YEAR = 2012

)

head(population\_data\_by\_year\_2012,5)

population\_data\_by\_year\_2013 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2013,

FIRE\_YEAR = 2013

)

population\_data\_by\_year\_2014 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2014,

FIRE\_YEAR = 2014

)

population\_data\_by\_year\_2015 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2015,

FIRE\_YEAR = 2015

)

population\_data\_by\_year\_2016 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2016,

FIRE\_YEAR = 2016

)

population\_data\_by\_year\_2017 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2017,

FIRE\_YEAR = 2017

)

population\_data\_by\_year\_2018 <- data.frame(

COUNTY = df$County,

STATE = df$State,

POPULATION = df$Year\_2018,

FIRE\_YEAR = 2018

)

population\_data\_by\_year <- rbind(population\_data\_by\_year\_2010, population\_data\_by\_year\_2011,

population\_data\_by\_year\_2012, population\_data\_by\_year\_2013,

population\_data\_by\_year\_2014, population\_data\_by\_year\_2015,

population\_data\_by\_year\_2016, population\_data\_by\_year\_2017,

population\_data\_by\_year\_2018)

fileName <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level", "population\_by\_year.csv")

ret <- write.csv(x=population\_data\_by\_year, fileName)

print(ret)

fire\_data\_with\_population <- merge(population\_data\_by\_year, cleanup\_forest\_fire\_data,by=c("STATE","COUNTY","FIRE\_YEAR"))

fileName <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level", "fire\_data\_with\_population.csv")

ret <- write.csv(x=fire\_data\_with\_population, fileName)

population\_census <- import("Poverty\_2010\_Census.xlsx")

head(population\_census,5)

fire\_data\_with\_population\_with\_poverty <- merge(fire\_data\_with\_population, population\_census,by=c("STATE","COUNTY"))

head(fire\_data\_with\_population\_with\_poverty,5)

fire\_data\_with\_population\_with\_poverty\_file <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level",

"fire\_data\_with\_population\_with\_poverty.csv")

ret <- write.csv(x=fire\_data\_with\_population\_with\_poverty, fire\_data\_with\_population\_with\_poverty\_file)

housing\_data <- import("Housing\_2010-19\_clean.xlsx")

housing\_data\_by\_year\_2010 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2010,

FIRE\_YEAR = 2010

)

housing\_data\_by\_year\_2011 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2011,

FIRE\_YEAR = 2011

)

housing\_data\_by\_year\_2012 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2012,

FIRE\_YEAR = 2012

)

housing\_data\_by\_year\_2013 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2013,

FIRE\_YEAR = 2013

)

housing\_data\_by\_year\_2014 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2014,

FIRE\_YEAR = 2014

)

housing\_data\_by\_year\_2015 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2015,

FIRE\_YEAR = 2015

)

housing\_data\_by\_year\_2016 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2016,

FIRE\_YEAR = 2016

)

housing\_data\_by\_year\_2017 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2017,

FIRE\_YEAR = 2017

)

housing\_data\_by\_year\_2018 <- data.frame(

COUNTY = housing\_data$COUNTY,

STATE = housing\_data$STATE,

CENSUS = housing\_data$Census,

ESTIMATES\_BASE = housing\_data$`Estimates Base`,

HOUSING = housing\_data$Year\_2018,

FIRE\_YEAR = 2018

)

housing\_data\_by\_year <- rbind(housing\_data\_by\_year\_2010, housing\_data\_by\_year\_2011,housing\_data\_by\_year\_2012,

housing\_data\_by\_year\_2013,housing\_data\_by\_year\_2014,housing\_data\_by\_year\_2015,

housing\_data\_by\_year\_2016,housing\_data\_by\_year\_2017,housing\_data\_by\_year\_2018)

fire\_data\_with\_population\_with\_poverty\_housing <- merge(housing\_data\_by\_year, fire\_data\_with\_population\_with\_poverty, by=c("STATE","COUNTY","FIRE\_YEAR"))

fire\_data\_with\_population\_with\_poverty\_file <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level",

"fire\_data\_with\_population\_with\_poverty\_with\_housing.csv")

ret <- write.csv(x=fire\_data\_with\_population\_with\_poverty\_housing, fire\_data\_with\_population\_with\_poverty\_file)

unemployment\_count <- import("Unemployment\_count\_2010-20.xlsx")

head(unemployment\_count,5)

unemployment\_count\_by\_year\_2010 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2010,

FIRE\_YEAR = 2010

)

unemployment\_count\_by\_year\_2011 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2011,

FIRE\_YEAR = 2011

)

unemployment\_count\_by\_year\_2012 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2012,

FIRE\_YEAR = 2012

)

unemployment\_count\_by\_year\_2013 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2013,

FIRE\_YEAR = 2013

)

unemployment\_count\_by\_year\_2014 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2014,

FIRE\_YEAR = 2014

)

unemployment\_count\_by\_year\_2015 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2015,

FIRE\_YEAR = 2015

)

unemployment\_count\_by\_year\_2016 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2016,

FIRE\_YEAR = 2016

)

unemployment\_count\_by\_year\_2017 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2017,

FIRE\_YEAR = 2017

)

unemployment\_count\_by\_year\_2018 <- data.frame(

COUNTY = unemployment\_count$COUNTY,

STATE = unemployment\_count$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_count$Year\_2018,

FIRE\_YEAR = 2018

)

unemployment\_count\_by\_year <- rbind(unemployment\_count\_by\_year\_2010,unemployment\_count\_by\_year\_2011,unemployment\_count\_by\_year\_2012,

unemployment\_count\_by\_year\_2013,unemployment\_count\_by\_year\_2014,unemployment\_count\_by\_year\_2015,

unemployment\_count\_by\_year\_2016,unemployment\_count\_by\_year\_2017,unemployment\_count\_by\_year\_2018)

fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt <- merge(fire\_data\_with\_population\_with\_poverty\_housing, unemployment\_count\_by\_year, by=c("STATE","COUNTY","FIRE\_YEAR"))

fire\_data\_with\_population\_poverty\_housing\_unemployment\_file <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level",

"fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt.csv")

ret <- write.csv(x=fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt,fire\_data\_with\_population\_poverty\_housing\_unemployment\_file)

unemployment\_rate <- import("Unemployment\_rate\_2010-20.xlsx")

head(unemployment\_rate,5)

unemployment\_rate\_by\_year\_2010 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2010,

FIRE\_YEAR = 2010

)

unemployment\_rate\_by\_year\_2011 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2011,

FIRE\_YEAR = 2011

)

unemployment\_rate\_by\_year\_2012 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2012,

FIRE\_YEAR = 2012

)

unemployment\_rate\_by\_year\_2013 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2013,

FIRE\_YEAR = 2013

)

unemployment\_rate\_by\_year\_2014 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2014,

FIRE\_YEAR = 2014

)

unemployment\_rate\_by\_year\_2015 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2015,

FIRE\_YEAR = 2015

)

unemployment\_rate\_by\_year\_2016 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2016,

FIRE\_YEAR = 2016

)

unemployment\_rate\_by\_year\_2017 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2017,

FIRE\_YEAR = 2017

)

unemployment\_rate\_by\_year\_2018 <- data.frame(

COUNTY = unemployment\_rate$COUNTY,

STATE = unemployment\_rate$STATE,

UNEMPLOYMENT\_COUNT = unemployment\_rate$Year\_2018,

FIRE\_YEAR = 2018

)

unemployment\_rate\_by\_year <- rbind(unemployment\_rate\_by\_year\_2010,unemployment\_rate\_by\_year\_2011,

unemployment\_rate\_by\_year\_2012,unemployment\_rate\_by\_year\_2013,

unemployment\_rate\_by\_year\_2014,unemployment\_rate\_by\_year\_2015,

unemployment\_rate\_by\_year\_2016,unemployment\_rate\_by\_year\_2017,

unemployment\_rate\_by\_year\_2018)

head(unemployment\_rate\_by\_year,5)

fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt\_rate <-

merge(fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt,unemployment\_rate\_by\_year, by=c("STATE","COUNTY","FIRE\_YEAR"))

fire\_data\_with\_population\_poverty\_housing\_unemployment\_file <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level",

"fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt\_rate.csv")

ret <- write.csv(x=fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt\_rate, fire\_data\_with\_population\_poverty\_housing\_unemployment\_file)

print(ret)

median\_income <- import("Median Household Income\_2014-18.xlsx")

head(median\_income,5)

median\_income\_data <- data.frame(

STATE = median\_income$STATE,

COUNTY = toupper(median\_income$COUNTY),

median\_income = median\_income$`Median Household Income (2014-18)`

)

fire\_data\_population\_poverty\_housing\_unemployment\_cnt\_rate\_median\_income <- merge(fire\_data\_with\_population\_poverty\_housing\_unemployment\_cnt\_rate, median\_income\_data, by=c("STATE","COUNTY"))

fire\_data\_population\_poverty\_housing\_unemployment\_cnt\_rate\_median\_income\_file <- file.path("/Users/abhinitkumarsinha/Documents/Surbhi\_FinalYear\_Project/FINAL DATA FILES/Clean Data/Socio-economic data County Level",

"fire\_data\_population\_poverty\_housing\_unemployment\_cnt\_rate\_median\_income.csv")

ret <- write.csv(x=fire\_data\_population\_poverty\_housing\_unemployment\_cnt\_rate\_median\_income, fire\_data\_population\_poverty\_housing\_unemployment\_cnt\_rate\_median\_income\_file)

***Naïve Bayes***

*# DATA PREPARATION*

*## Reading data*

DF <- read.csv("DF.csv")

dim(DF)

DF1 <- DF[!(DF$NWCG\_GENERAL\_CAUSE=="Missing data/not specified/undetermined"),] *# Dropped records of missing cause*

dim(DF1)

DF2 <- DF1[!(DF1$NWCG\_GENERAL\_CAUSE=="Other causes"),] *# Dropped other causes*

dim(DF2)

*## Converting Categorical Columns to numeric*

DF2$DISCOVERY\_MONTH <- match(DF2$DISCOVERY\_MONTH, month.name)

library(dplyr)

DF2$DISCOVERY\_DAY <- case\_when(DF2$DISCOVERY\_DAY == "Sunday" ~ '1',

DF2$DISCOVERY\_DAY == "Monday" ~ '2',

DF2$DISCOVERY\_DAY == "Tuesday" ~ '3',

DF2$DISCOVERY\_DAY == "Wednesday" ~ '4',

DF2$DISCOVERY\_DAY == "Thursday" ~ '5',

DF2$DISCOVERY\_DAY == "Friday" ~ '6',

DF2$DISCOVERY\_DAY == "Saturday" ~ '7')

DF2$SEASON <- case\_when(DF2$SEASON == "Spring" ~'1',

DF2$SEASON == "Summer" ~'2',

DF2$SEASON == "Fall" ~'3',

DF2$SEASON == "Winter" ~'4')

DF2$FIRE\_SIZE\_CLASS <- case\_when(DF2$FIRE\_SIZE\_CLASS =="A"~'1',

DF2$FIRE\_SIZE\_CLASS =="B"~'2',

DF2$FIRE\_SIZE\_CLASS =="C"~'3',

DF2$FIRE\_SIZE\_CLASS =="D"~'4',

DF2$FIRE\_SIZE\_CLASS =="E"~'5',

DF2$FIRE\_SIZE\_CLASS =="F"~'6',

DF2$FIRE\_SIZE\_CLASS =="G"~'7')

*## Creating factors for Response Variable*

unique(DF2$NWCG\_GENERAL\_CAUSE)

DF2$NWCG\_GENERAL\_CAUSE <- as.integer(factor(DF2$NWCG\_GENERAL\_CAUSE,levels=c("Power generation/transmission/distribution",

"Debris and open burning",

"Misuse of fire by a minor",

"Smoking",

"Arson/incendiarism",

"Equipment and vehicle use",

"Recreation and ceremony",

"Fireworks",

"Railroad operations and maintenance",

"Firearms and explosives use")

,ordered = **TRUE**))

*## Dropping descriptive columns*

DF3 <- DF2[,-c(1,2,5,6,11)]

*## Finding Correlation between numeric columns*

temp <- DF3[,c(10:16)]

library(corrplot)

corrplot(cor(temp),method = "number", type = "upper")

library(caret)

findCorrelation(cor(temp),cutoff = 0.8, verbose = **TRUE**, names = **TRUE**, exact = **FALSE**)

*## Partitioning into 3 sets: Training (60%), Validation (20%), Testing (20%)*

set.seed(1)

total.rows <- nrow(DF3)

total.rows

train.size <- floor(.6\*total.rows)

validation.size <- floor(.2\*total.rows)

train.rows <- sample(1:total.rows, train.size)

training.set <- DF3[train.rows,]

remaining.set <- DF3[-train.rows,]

remaining.rows <- nrow(remaining.set)

validation.rows <- sample(1:remaining.rows,validation.size)

validation.set <- remaining.set[validation.rows,]

testing.set <- remaining.set[-validation.rows,]

dim(training.set)

dim(validation.set)

dim(testing.set)

training.set$NWCG\_GENERAL\_CAUSE <- as.factor(training.set$NWCG\_GENERAL\_CAUSE)

validation.set$NWCG\_GENERAL\_CAUSE <- as.factor(validation.set$NWCG\_GENERAL\_CAUSE)

testing.set$NWCG\_GENERAL\_CAUSE <- as.factor(testing.set$NWCG\_GENERAL\_CAUSE)

*# APPLICATION OF ML ALGORITHMS*

*## Installing required packages*

**if** (!require("pacman")) install.packages("pacman","sos","xlsx","kernlab","e1071","naivebayes","adabag", "klaR")

pacman::p\_load(pacman,rio,tidyverse,sos,kernlab,e1071,naivebayes,adabag)

library(readxl)

library(naivebayes)

library(dplyr)

library(adabag)

library(rpart.plot)

library(party)

library(pROC)

library(klaR)

*## Naive Bayes*

*### Creating model with highly correlated columns*

model1 <- NaiveBayes(NWCG\_GENERAL\_CAUSE ~ ., data=training.set)

test.pred1 <- predict(model1, testing.set)

roc\_multiclass1 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test.pred1$posterior)

auc(roc\_multiclass1) *# 0.6303*

*## Creating model without highly correlated columns*

model2 <- NaiveBayes(NWCG\_GENERAL\_CAUSE ~ ., data=training.set[,-c(10,11,13)])

test.pred2 <- predict(model2, testing.set[,-c(10,11,13)])

roc\_multiclass2 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test.pred2$posterior)

auc(roc\_multiclass2) *# 0.6517*

*## Multi-nomial Logistic Regression*

***Multinomial Logistic Regression***

*# DATA PREPARATION*

*## Reading data*

DF <- read.csv("DF.csv")

dim(DF)

DF1 <- DF[!(DF$NWCG\_GENERAL\_CAUSE=="Missing data/not specified/undetermined"),] *# Dropped records of missing cause*

dim(DF1)

DF2 <- DF1[!(DF1$NWCG\_GENERAL\_CAUSE=="Other causes"),] *# Dropped other causes*

dim(DF2)

*## Converting Categorical Columns to numeric*

DF2$DISCOVERY\_MONTH <- match(DF2$DISCOVERY\_MONTH, month.name)

library(dplyr)

DF2$DISCOVERY\_DAY <- case\_when(DF2$DISCOVERY\_DAY == "Sunday" ~ '1',

DF2$DISCOVERY\_DAY == "Monday" ~ '2',

DF2$DISCOVERY\_DAY == "Tuesday" ~ '3',

DF2$DISCOVERY\_DAY == "Wednesday" ~ '4',

DF2$DISCOVERY\_DAY == "Thursday" ~ '5',

DF2$DISCOVERY\_DAY == "Friday" ~ '6',

DF2$DISCOVERY\_DAY == "Saturday" ~ '7')

DF2$SEASON <- case\_when(DF2$SEASON == "Spring" ~'1',

DF2$SEASON == "Summer" ~'2',

DF2$SEASON == "Fall" ~'3',

DF2$SEASON == "Winter" ~'4')

DF2$FIRE\_SIZE\_CLASS <- case\_when(DF2$FIRE\_SIZE\_CLASS =="A"~'1',

DF2$FIRE\_SIZE\_CLASS =="B"~'2',

DF2$FIRE\_SIZE\_CLASS =="C"~'3',

DF2$FIRE\_SIZE\_CLASS =="D"~'4',

DF2$FIRE\_SIZE\_CLASS =="E"~'5',

DF2$FIRE\_SIZE\_CLASS =="F"~'6',

DF2$FIRE\_SIZE\_CLASS =="G"~'7')

*## Creating factors for Response Variable*

unique(DF2$NWCG\_GENERAL\_CAUSE)

DF2$NWCG\_GENERAL\_CAUSE <- as.integer(factor(DF2$NWCG\_GENERAL\_CAUSE,levels=c("Power generation/transmission/distribution",

"Debris and open burning",

"Misuse of fire by a minor",

"Smoking",

"Arson/incendiarism",

"Equipment and vehicle use",

"Recreation and ceremony",

"Fireworks",

"Railroad operations and maintenance",

"Firearms and explosives use")

,ordered = **TRUE**))

*## Dropping descriptive columns*

DF3 <- DF2[,-c(1,2,5,6,11)]

*## Finding Correlation between numeric columns*

temp <- DF3[,c(10:16)]

library(corrplot)

corrplot(cor(temp),method = "number", type = "upper")

library(caret)

findCorrelation(cor(temp),cutoff = 0.8, verbose = **TRUE**, names = **TRUE**, exact = **FALSE**)

*## Partitioning into 3 sets: Training (60%), Validation (20%), Testing (20%)*

set.seed(1)

total.rows <- nrow(DF3)

total.rows

train.size <- floor(.6\*total.rows)

validation.size <- floor(.2\*total.rows)

train.rows <- sample(1:total.rows, train.size)

training.set <- DF3[train.rows,]

remaining.set <- DF3[-train.rows,]

remaining.rows <- nrow(remaining.set)

validation.rows <- sample(1:remaining.rows,validation.size)

validation.set <- remaining.set[validation.rows,]

testing.set <- remaining.set[-validation.rows,]

dim(training.set)

dim(validation.set)

dim(testing.set)

training.set$NWCG\_GENERAL\_CAUSE <- as.factor(training.set$NWCG\_GENERAL\_CAUSE)

validation.set$NWCG\_GENERAL\_CAUSE <- as.factor(validation.set$NWCG\_GENERAL\_CAUSE)

testing.set$NWCG\_GENERAL\_CAUSE <- as.factor(testing.set$NWCG\_GENERAL\_CAUSE)

*# APPLICATION OF ML ALGORITHMS*

*## Installing required packages*

**if** (!require("pacman")) install.packages("pacman","sos","xlsx","kernlab","e1071","naivebayes","adabag", "klaR")

pacman::p\_load(pacman,rio,tidyverse,sos,kernlab,e1071,naivebayes,adabag)

library(readxl)

library(naivebayes)

library(dplyr)

library(adabag)

library(rpart.plot)

library(party)

library(pROC)

library(klaR)

*## Multinomial Logistic Regression*

*### Creating model with highly correlated columns*

model <- nnet::multinom(NWCG\_GENERAL\_CAUSE ~ ., data=training.set)

pp <- fitted(model)

pp

validation\_predicted <- predict(model, validation.set)

tab <- table(validation\_predicted, validation.set$NWCG\_GENERAL\_CAUSE)

sum(diag(tab))/sum(tab)

model\_summary <- summary(model)

z <- model\_summary$coefficients/model\_summary$standard.errors

p <- (1 - pnorm(abs(z),0,1))\*2

p

multinomial\_logistic\_reg <- nnet::multinom(NWCG\_GENERAL\_CAUSE ~ ., data=training.set)

test\_data\_predicted <- predict(multinomial\_logistic\_reg, testing.set)

multinom\_test\_data\_predicted\_prob <- predict(multinomial\_logistic\_reg, testing.set, type='prob')

multinomial\_logistic\_roc\_multiclass <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE,multinom\_test\_data\_predicted\_prob)

auc(multinomial\_logistic\_roc\_multiclass)

*### Creating model without highly correlated columns*

multinomial\_logistic\_reg2 <- nnet::multinom(NWCG\_GENERAL\_CAUSE ~ ., data=training.set[,-c(10,11,13)])

test\_data\_predicted2 <- predict(multinomial\_logistic\_reg2, testing.set[,-c(10,11,13)])

multinom\_test\_data\_predicted\_prob2 <- predict(multinomial\_logistic\_reg2, testing.set[,-c(10,11,13)], type='prob')

multinomial\_logistic\_roc\_multiclass2 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE,multinom\_test\_data\_predicted\_prob2)

auc(multinomial\_logistic\_roc\_multiclass2)

***Random Forests***

*# DATA PREPARATION*

*## Reading data*

DF <- read.csv("DF.csv")

dim(DF)

DF1 <- DF[!(DF$NWCG\_GENERAL\_CAUSE=="Missing data/not specified/undetermined"),] *# Dropped records of missing cause*

dim(DF1)

DF2 <- DF1[!(DF1$NWCG\_GENERAL\_CAUSE=="Other causes"),] *# Dropped other causes*

dim(DF2)

*## Converting Categorical Columns to numeric*

DF2$DISCOVERY\_MONTH <- match(DF2$DISCOVERY\_MONTH, month.name)

library(dplyr)

DF2$DISCOVERY\_DAY <- case\_when(DF2$DISCOVERY\_DAY == "Sunday" ~ '1',

DF2$DISCOVERY\_DAY == "Monday" ~ '2',

DF2$DISCOVERY\_DAY == "Tuesday" ~ '3',

DF2$DISCOVERY\_DAY == "Wednesday" ~ '4',

DF2$DISCOVERY\_DAY == "Thursday" ~ '5',

DF2$DISCOVERY\_DAY == "Friday" ~ '6',

DF2$DISCOVERY\_DAY == "Saturday" ~ '7')

DF2$SEASON <- case\_when(DF2$SEASON == "Spring" ~'1',

DF2$SEASON == "Summer" ~'2',

DF2$SEASON == "Fall" ~'3',

DF2$SEASON == "Winter" ~'4')

DF2$FIRE\_SIZE\_CLASS <- case\_when(DF2$FIRE\_SIZE\_CLASS =="A"~'1',

DF2$FIRE\_SIZE\_CLASS =="B"~'2',

DF2$FIRE\_SIZE\_CLASS =="C"~'3',

DF2$FIRE\_SIZE\_CLASS =="D"~'4',

DF2$FIRE\_SIZE\_CLASS =="E"~'5',

DF2$FIRE\_SIZE\_CLASS =="F"~'6',

DF2$FIRE\_SIZE\_CLASS =="G"~'7')

*## Creating factors for Response Variable*

unique(DF2$NWCG\_GENERAL\_CAUSE)

DF2$NWCG\_GENERAL\_CAUSE <- as.integer(factor(DF2$NWCG\_GENERAL\_CAUSE,levels=c("Power generation/transmission/distribution",

"Debris and open burning",

"Misuse of fire by a minor",

"Smoking",

"Arson/incendiarism",

"Equipment and vehicle use",

"Recreation and ceremony",

"Fireworks",

"Railroad operations and maintenance",

"Firearms and explosives use")

,ordered = **TRUE**))

*## Dropping descriptive columns*

DF3 <- DF2[,-c(1,2,5,6,11)]

*## Finding Correlation between numeric columns*

temp <- DF3[,c(10:16)]

library(corrplot)

corrplot(cor(temp),method = "number", type = "upper")

library(caret)

findCorrelation(cor(temp),cutoff = 0.8, verbose = **TRUE**, names = **TRUE**, exact = **FALSE**)

*## Partitioning into 3 sets: Training (60%), Validation (20%), Testing (20%)*

set.seed(1)

total.rows <- nrow(DF3)

total.rows

train.size <- floor(.6\*total.rows)

validation.size <- floor(.2\*total.rows)

train.rows <- sample(1:total.rows, train.size)

training.set <- DF3[train.rows,]

remaining.set <- DF3[-train.rows,]

remaining.rows <- nrow(remaining.set)

validation.rows <- sample(1:remaining.rows,validation.size)

validation.set <- remaining.set[validation.rows,]

testing.set <- remaining.set[-validation.rows,]

dim(training.set)

dim(validation.set)

dim(testing.set)

training.set$NWCG\_GENERAL\_CAUSE <- as.factor(training.set$NWCG\_GENERAL\_CAUSE)

validation.set$NWCG\_GENERAL\_CAUSE <- as.factor(validation.set$NWCG\_GENERAL\_CAUSE)

testing.set$NWCG\_GENERAL\_CAUSE <- as.factor(testing.set$NWCG\_GENERAL\_CAUSE)

*# APPLICATION OF ML ALGORITHMS*

*## Installing required packages*

**if** (!require("pacman")) install.packages("pacman","sos","xlsx","kernlab","e1071","naivebayes","adabag", "klaR")

pacman::p\_load(pacman,rio,tidyverse,sos,kernlab,e1071,naivebayes,adabag)

library(readxl)

library(naivebayes)

library(dplyr)

library(adabag)

library(rpart.plot)

library(party)

library(pROC)

library(klaR)

*## Random Forests*

library(randomForest)

rf <- randomForest(NWCG\_GENERAL\_CAUSE ~ ., data=training.set, ntree=200, mtry=6,importance=**TRUE**)

plot(rf)

train\_prediction\_rf <- predict(rf, training\_set\_numeric)

confusion\_matrix\_rf <- confusionMatrix(train\_prediction\_rf, training\_set\_numeric$NWCG\_GENERAL\_CAUSE)

validation\_prediction\_rf <- predict(rf, validation\_set\_numeric)

cf\_matrix\_validation\_prediction <- confusionMatrix(validation\_prediction\_rf, validation\_set\_numeric$NWCG\_GENERAL\_CAUSE)

*#finds optimal value of mtry*

tuned\_params <- tuneRF(training.set[,-7],training.set[,7],

stepFactor = 0.5,

plot = **TRUE**,

ntreeTry = 200,

trace=**TRUE**,

improve = 5)

varImpPlot(rf)

test\_data\_predicted<- predict(rf, testing.set, type='prob')

roc\_multiclass\_auc <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test\_data\_predicted)

auc(roc\_multiclass\_auc)

*### Creating Model with highly correlated columns*

rf <- randomForest(NWCG\_GENERAL\_CAUSE ~ ., data=training.set, ntree=200, mtry=8,importance=**TRUE**)

*### Creating Model without highly correlated columns*

rf2 <- randomForest(NWCG\_GENERAL\_CAUSE ~ ., data=training.set[,-c(10,11,13)], ntree=200, mtry=6,importance=**TRUE**)

test\_data\_predicted2 <- predict(rf2, testing.set[,-c(10,11,13)], type='prob')

roc\_multiclass\_auc2 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test\_data\_predicted2)

auc(roc\_multiclass\_auc2)

***Boosting***

*# DATA PREPARATION*

*## Reading data*

DF <- read.csv("DF.csv")

dim(DF)

DF1 <- DF[!(DF$NWCG\_GENERAL\_CAUSE=="Missing data/not specified/undetermined"),] *# Dropped records of missing cause*

dim(DF1)

DF2 <- DF1[!(DF1$NWCG\_GENERAL\_CAUSE=="Other causes"),] *# Dropped other causes*

dim(DF2)

*## Converting Categorical Columns to numeric*

DF2$DISCOVERY\_MONTH <- match(DF2$DISCOVERY\_MONTH, month.name)

library(dplyr)

DF2$DISCOVERY\_DAY <- case\_when(DF2$DISCOVERY\_DAY == "Sunday" ~ '1',

DF2$DISCOVERY\_DAY == "Monday" ~ '2',

DF2$DISCOVERY\_DAY == "Tuesday" ~ '3',

DF2$DISCOVERY\_DAY == "Wednesday" ~ '4',

DF2$DISCOVERY\_DAY == "Thursday" ~ '5',

DF2$DISCOVERY\_DAY == "Friday" ~ '6',

DF2$DISCOVERY\_DAY == "Saturday" ~ '7')

DF2$SEASON <- case\_when(DF2$SEASON == "Spring" ~'1',

DF2$SEASON == "Summer" ~'2',

DF2$SEASON == "Fall" ~'3',

DF2$SEASON == "Winter" ~'4')

DF2$FIRE\_SIZE\_CLASS <- case\_when(DF2$FIRE\_SIZE\_CLASS =="A"~'1',

DF2$FIRE\_SIZE\_CLASS =="B"~'2',

DF2$FIRE\_SIZE\_CLASS =="C"~'3',

DF2$FIRE\_SIZE\_CLASS =="D"~'4',

DF2$FIRE\_SIZE\_CLASS =="E"~'5',

DF2$FIRE\_SIZE\_CLASS =="F"~'6',

DF2$FIRE\_SIZE\_CLASS =="G"~'7')

*## Creating factors for Response Variable*

unique(DF2$NWCG\_GENERAL\_CAUSE)

DF2$NWCG\_GENERAL\_CAUSE <- as.integer(factor(DF2$NWCG\_GENERAL\_CAUSE,levels=c("Power generation/transmission/distribution",

"Debris and open burning",

"Misuse of fire by a minor",

"Smoking",

"Arson/incendiarism",

"Equipment and vehicle use",

"Recreation and ceremony",

"Fireworks",

"Railroad operations and maintenance",

"Firearms and explosives use")

,ordered = **TRUE**))

*## Dropping descriptive columns*

DF3 <- DF2[,-c(1,2,5,6,11)]

*## Finding Correlation between numeric columns*

temp <- DF3[,c(10:16)]

library(corrplot)

corrplot(cor(temp),method = "number", type = "upper")

library(caret)

findCorrelation(cor(temp),cutoff = 0.8, verbose = **TRUE**, names = **TRUE**, exact = **FALSE**)

*## Partitioning into 3 sets: Training (60%), Validation (20%), Testing (20%)*

set.seed(1)

total.rows <- nrow(DF3)

total.rows

train.size <- floor(.6\*total.rows)

validation.size <- floor(.2\*total.rows)

train.rows <- sample(1:total.rows, train.size)

training.set <- DF3[train.rows,]

remaining.set <- DF3[-train.rows,]

remaining.rows <- nrow(remaining.set)

validation.rows <- sample(1:remaining.rows,validation.size)

validation.set <- remaining.set[validation.rows,]

testing.set <- remaining.set[-validation.rows,]

dim(training.set)

dim(validation.set)

dim(testing.set)

training.set$NWCG\_GENERAL\_CAUSE <- as.factor(training.set$NWCG\_GENERAL\_CAUSE)

validation.set$NWCG\_GENERAL\_CAUSE <- as.factor(validation.set$NWCG\_GENERAL\_CAUSE)

testing.set$NWCG\_GENERAL\_CAUSE <- as.factor(testing.set$NWCG\_GENERAL\_CAUSE)

*# APPLICATION OF ML ALGORITHMS*

*## Installing required packages*

**if** (!require("pacman")) install.packages("pacman","sos","xlsx","kernlab","e1071","naivebayes","adabag", "klaR")

pacman::p\_load(pacman,rio,tidyverse,sos,kernlab,e1071,naivebayes,adabag)

library(readxl)

library(naivebayes)

library(dplyr)

library(adabag)

library(rpart.plot)

library(party)

library(pROC)

library(klaR)

*## Boosting*

*### Creating Model with highly correlated columns*

library(randomForest)

*# Boosting*

modelLookup("xgbTree")

*# cross validation with repetition*

cv <- trainControl(method = 'repeatedcv', number = 10, repeats = 5, allowParallel = **TRUE**)

*# training Xg Boost model with training data set and repeated cross validation*

boo <- train(NWCG\_GENERAL\_CAUSE ~ .,

data=training.set,

method="xgbTree",

trControl=cv,

tuneGrid = expand.grid(nrounds = 500,

max\_depth = 3,

eta = 0.3,

gamma = 0.01,

colsample\_bytree = 1,

min\_child\_weight = 1,

subsample = 1))

*#p1 <- predict(boo, newdata=validation.set, type = 'raw')*

p <- predict(boo, newdata=validation.set, type = 'prob')

test\_set\_prediction <- predict(boo, newdata=testing.set, type = 'prob')

roc\_multiclass <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test\_set\_prediction)

*#Area under Curve(AUC)*

auc(roc\_multiclass)

*### Creating Model without highly correlated columns*

boo1 <- train(NWCG\_GENERAL\_CAUSE ~ .,

data=training.set[,-c(10,11,13)],

method="xgbTree",

trControl=cv,

tuneGrid = expand.grid(nrounds = 500,

max\_depth = 3,

eta = 0.3,

gamma = 0.01,

colsample\_bytree = 1,

min\_child\_weight = 1,

subsample = 1))

*#p2 <- predict(boo1, newdata=validation.set, type = 'prob')*

test\_set\_prediction2 <- predict(boo1, newdata=testing.set[,-c(10,11,13)], type = 'prob')

roc\_multiclass2 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, test\_set\_prediction2)

*#Area under Curve(AUC)*

auc(roc\_multiclass2)

***KNN***

*# DATA PREPARATION*

*## Reading data*

DF <- read.csv("DF.csv")

dim(DF)

DF1 <- DF[!(DF$NWCG\_GENERAL\_CAUSE=="Missing data/not specified/undetermined"),] *# Dropped records of missing cause*

dim(DF1)

DF2 <- DF1[!(DF1$NWCG\_GENERAL\_CAUSE=="Other causes"),] *# Dropped other causes*

dim(DF2)

*## Converting Categorical Columns to numeric*

DF2$DISCOVERY\_MONTH <- match(DF2$DISCOVERY\_MONTH, month.name)

library(dplyr)

DF2$DISCOVERY\_DAY <- case\_when(DF2$DISCOVERY\_DAY == "Sunday" ~ '1',

DF2$DISCOVERY\_DAY == "Monday" ~ '2',

DF2$DISCOVERY\_DAY == "Tuesday" ~ '3',

DF2$DISCOVERY\_DAY == "Wednesday" ~ '4',

DF2$DISCOVERY\_DAY == "Thursday" ~ '5',

DF2$DISCOVERY\_DAY == "Friday" ~ '6',

DF2$DISCOVERY\_DAY == "Saturday" ~ '7')

DF2$SEASON <- case\_when(DF2$SEASON == "Spring" ~'1',

DF2$SEASON == "Summer" ~'2',

DF2$SEASON == "Fall" ~'3',

DF2$SEASON == "Winter" ~'4')

DF2$FIRE\_SIZE\_CLASS <- case\_when(DF2$FIRE\_SIZE\_CLASS =="A"~'1',

DF2$FIRE\_SIZE\_CLASS =="B"~'2',

DF2$FIRE\_SIZE\_CLASS =="C"~'3',

DF2$FIRE\_SIZE\_CLASS =="D"~'4',

DF2$FIRE\_SIZE\_CLASS =="E"~'5',

DF2$FIRE\_SIZE\_CLASS =="F"~'6',

DF2$FIRE\_SIZE\_CLASS =="G"~'7')

*## Creating factors for Response Variable*

unique(DF2$NWCG\_GENERAL\_CAUSE)

DF2$NWCG\_GENERAL\_CAUSE <- as.integer(factor(DF2$NWCG\_GENERAL\_CAUSE,levels=c("Power generation/transmission/distribution",

"Debris and open burning",

"Misuse of fire by a minor",

"Smoking",

"Arson/incendiarism",

"Equipment and vehicle use",

"Recreation and ceremony",

"Fireworks",

"Railroad operations and maintenance",

"Firearms and explosives use")

,ordered = **TRUE**))

*## Dropping descriptive columns*

DF3 <- DF2[,-c(1,2,5,6,11)]

*## Finding Correlation between numeric columns*

temp <- DF3[,c(10:16)]

library(corrplot)

corrplot(cor(temp),method = "number", type = "upper")

library(caret)

findCorrelation(cor(temp),cutoff = 0.8, verbose = **TRUE**, names = **TRUE**, exact = **FALSE**)

*## Partitioning into 3 sets: Training (60%), Validation (20%), Testing (20%)*

set.seed(1)

total.rows <- nrow(DF3)

total.rows

train.size <- floor(.6\*total.rows)

validation.size <- floor(.2\*total.rows)

train.rows <- sample(1:total.rows, train.size)

training.set <- DF3[train.rows,]

remaining.set <- DF3[-train.rows,]

remaining.rows <- nrow(remaining.set)

validation.rows <- sample(1:remaining.rows,validation.size)

validation.set <- remaining.set[validation.rows,]

testing.set <- remaining.set[-validation.rows,]

dim(training.set)

dim(validation.set)

dim(testing.set)

training.set$NWCG\_GENERAL\_CAUSE <- as.factor(training.set$NWCG\_GENERAL\_CAUSE)

validation.set$NWCG\_GENERAL\_CAUSE <- as.factor(validation.set$NWCG\_GENERAL\_CAUSE)

testing.set$NWCG\_GENERAL\_CAUSE <- as.factor(testing.set$NWCG\_GENERAL\_CAUSE)

*# APPLICATION OF ML ALGORITHMS*

*## Installing required packages*

**if** (!require("pacman")) install.packages("pacman","sos","xlsx","kernlab","e1071","naivebayes","adabag", "klaR")

pacman::p\_load(pacman,rio,tidyverse,sos,kernlab,e1071,naivebayes,adabag)

library(readxl)

library(naivebayes)

library(dplyr)

library(adabag)

library(rpart.plot)

library(party)

library(pROC)

library(klaR)

*## KNN*

*### Creating Model with highly correlated columns*

ctrl <- trainControl(method="repeatedcv",repeats = 3) *#,classProbs=TRUE,summaryFunction = twoClassSummary)*

knnFit <- train(NWCG\_GENERAL\_CAUSE ~ ., data = training.set, method = "knn", trControl = ctrl, preProcess = c("center","scale"), tuneLength = 20)

plot(knnFit)

predictions <- predict(knnFit, newdata = testing.set, type="prob")

knn\_roc\_multiclass <- multiclass.roc(test\_set\_numeric$NWCG\_GENERAL\_CAUSE, predictions)

auc(knn\_roc\_multiclass)

*### Creating Model without highly correlated columns*

knnFit2 <- train(NWCG\_GENERAL\_CAUSE ~ ., data = training.set[,-c(10,11,13)], method = "knn", trControl = ctrl, preProcess = c("center","scale"), tuneLength = 20)

plot(knnFit2)

predictions2 <- predict(knnFit2, newdata = testing.set[,-c(10,11,13)], type="prob")

knn\_roc\_multiclass2 <- multiclass.roc(testing.set$NWCG\_GENERAL\_CAUSE, predictions2)

auc(knn\_roc\_multiclass2)