Predicting the nature & extent of Forest Fires in New York State for Year 2015 & 2016

By Ankur Peshin | peshia@rpi.edu

Data Analytics Term Assignment Report

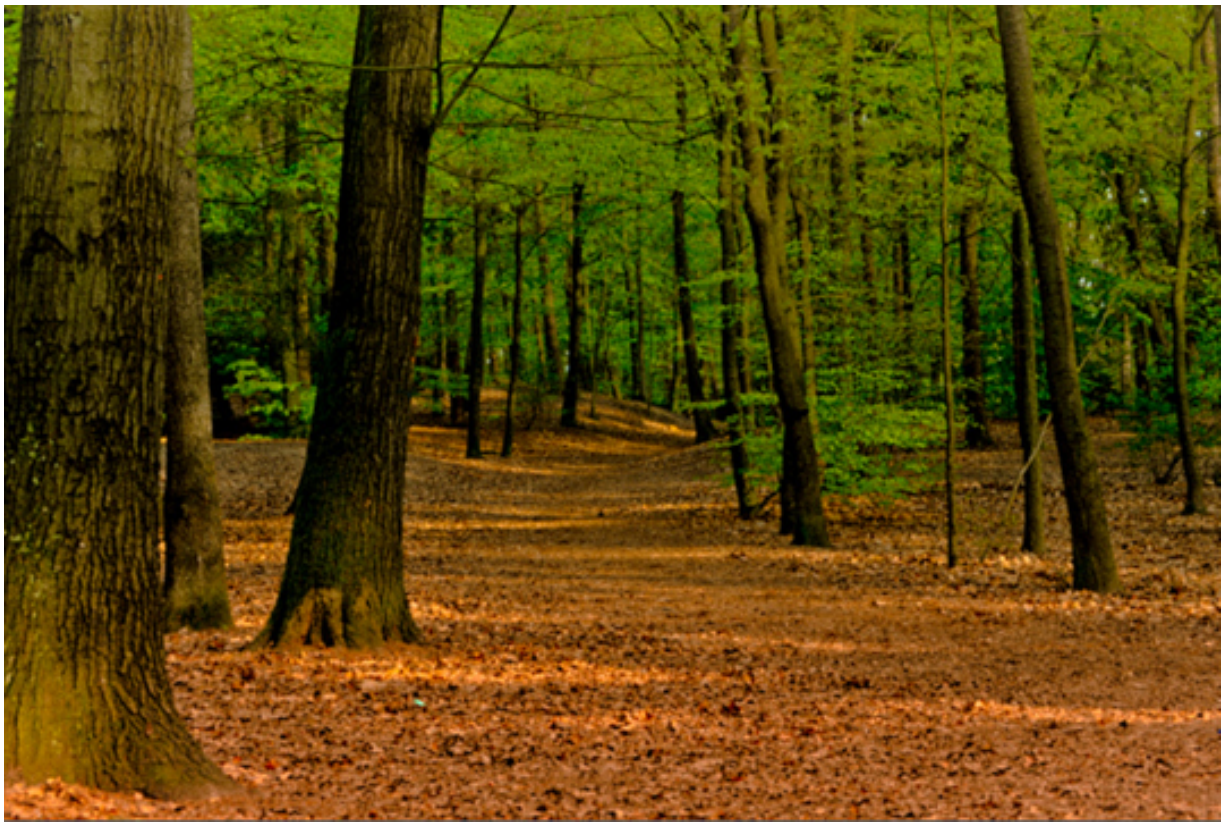


Table of Contents

Introduction

Wildfires, also known as forest fires, damage thousands of acres of natural resources every year in New York. Although wildfires naturally occur from lightning, the NYS Department of Environmental conservation has a different statistic which says that most forest fires are caused by human activities. However, not all forest fires are negative incidents. Prescribed fire is a method where forest rangers set controlled fire to the forest in order to create a artificial resilience of fire

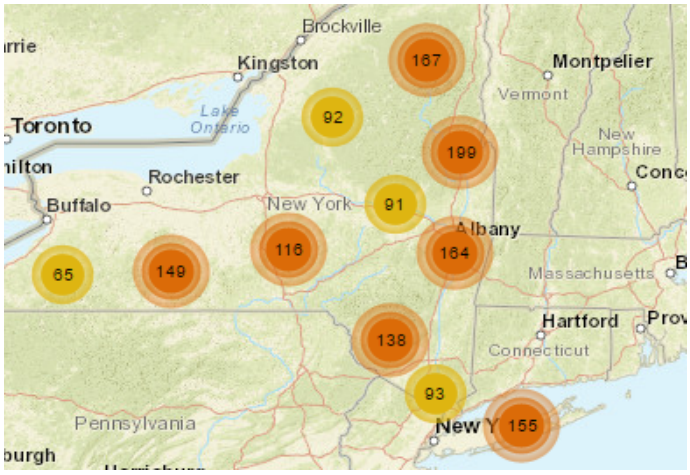
Prescribed fire is done during cooler months in order to reduce fuel buildup and prevent a likelihood of more serious forest fires. Prescribed fire also stimulates germination of desirable trees and renews vitality of soil.

Based on fire reports formulated by *Data.ny.gov,* which is a comprehensive report of Forest fire incidences starting year 2008 till year 2016 in New York State, I want to predict the acreage level burnt for 2015, by training the dataset from 2008-14, and acreage level burnt in 2016 by training the dataset from 2008-15. I also want to predict which fires were prescribed in 2015, by training my dataset from 2008-14, and which fires were prescribed in 2016, by training my dataset from 2008-15, using three different supervised learning approaches.

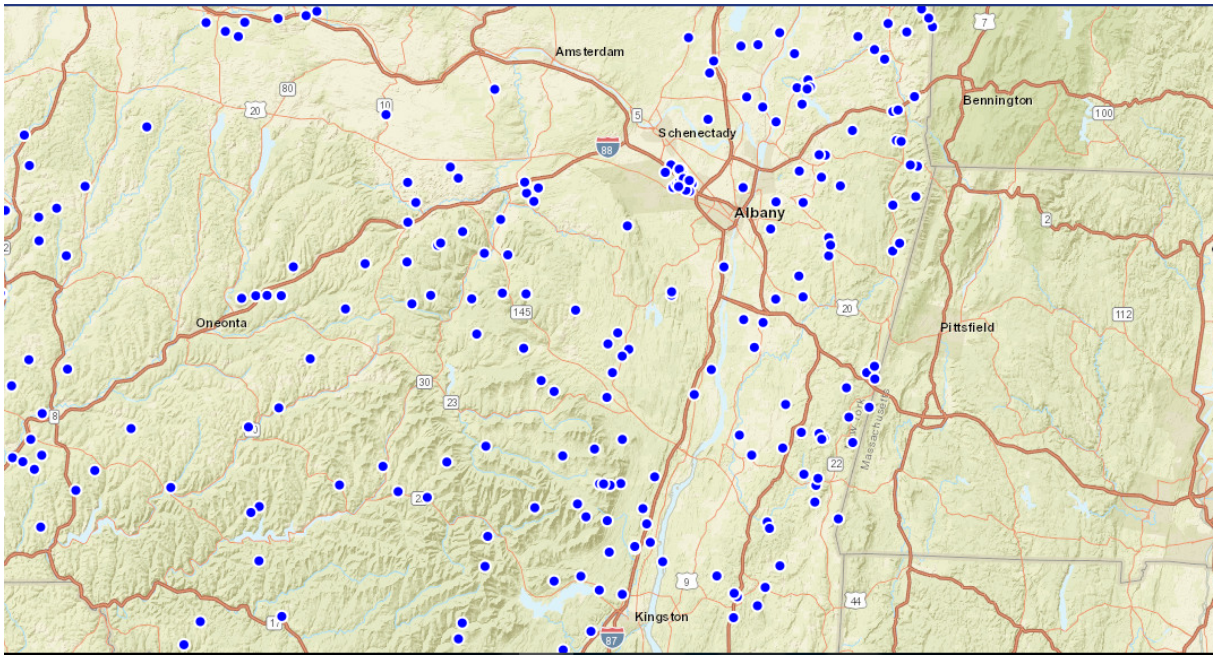
The motivation for pursuing this idea is the fact that I strongly advocate for environmental preservation, and I am a contributing member of Greenpeace India, therefore I was on the lookout of any kind of data which captures forest fires. Fortunately for me, the NYS Dept of Environmental Preservation has provided a very lucid dataset, with proper metadata information. I was glad that I did not have to do any row trims on accounts of outliers, or blank, null columns, for which I thank NYS Dept, because they did all the work related to Data Science due to which I could focus on Data Analytics.

Data Description

The dataset consists of 1430 rows and 28 columns of fire descriptions across all counties, regions , zones, and municipalities in NY state. The incident map of this dataset, which can be accessed from the references, takes the latitude and longitude information and plugs it into a map for visualization purposes.



The map also facilitates zoom-in feature, so that an interested person like a forest ranger can view the specificity of fire information and make decision based on that.



Although the dataset is very rich qualitatively, as seen by the number of columns, the quantity of data is somewhat less, which is a challenge worth accepting. There are no. of columns which are not causal to the variables I'm modelling. In other words, while trying to model the relation of dependant variable Y against X (Y~X), I have certain Y's on the left side of the equation which were necessary to weed out. Details of those are provided in Data Analysis Section.

Data Analysis

For the preliminary data Analysis, I had to clear off various columns which I thought would not be essential to the dependent variables I was going to predict. The information of columns removed are provided below along with Metadata information:

Fire Number	Unique index for wildland fires reported by NYS DEC Forest Rangers for national reporting.
Incident Name	Short narrative description of wildland fire that is unique for the current calendar year.
Railroad Name	Corporate name of the owner of the railroad property when the fire cause is classified "Railroad".
Location 1	Fire start location provided by DEC.
Fatalities	Number of people who were killed or died on the fire incident
Injuries	Number of people injured on the fire incident.
Homes Lost	Number of residential units that became uninhabitable by the fire

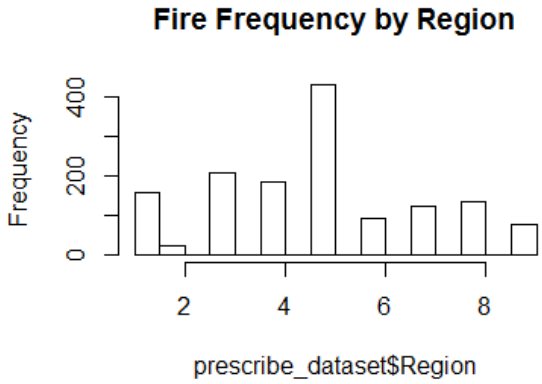
Homes Threatened	Number of residential units damaged or in threat of damage if fire suppression was not effective.
Other Structures Lost	Number of non-residential buildings severely damaged or destroyed by the fire.
Other Structures Threatened	Number of non-residential buildings damaged or in threat of damage if fire suppression was not effective.
Fire Report Method	Defined Category of what source the fire was reported to forest rangers.

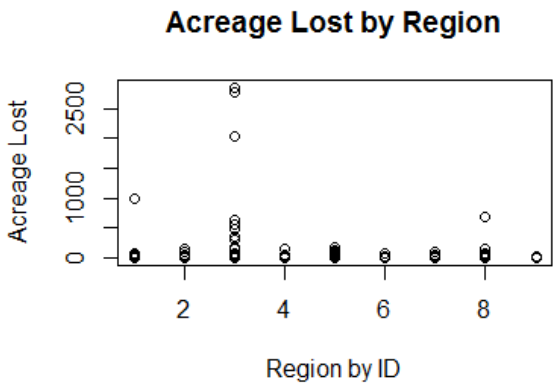
Justification for weeding these columns initially:

Columns	Justification
Fatalities, Injuries, Homes Lost,Homes Threatened, Other Structures Lost, Other Structures Threatened	Data was mostly zeros, with occasional non zero value so, correlation with Fire Cause or acreage will be low. Also these are causal effects of fire so they cannot be independent of Acreage, or cause.
Railroad Name	Mostly Blank columns, some with N/A, not precise to data Analytics
Location1	Split into latitude and latitude to individually study correlation.
Fire Number, Incident Name	These are continuous variables (like 1,2,3...so on), which will not produce results if taken into consideration
Reporting Ranger	Included this later, but did not help with prediction

After this operation, I was left with 18 columns of information, on which I started to do exploratory data Analysis. I named this data frame as *exploratory_df*, in order to study relations between various variables. I started first with examining the which counties have most incidences of fire using histogram.

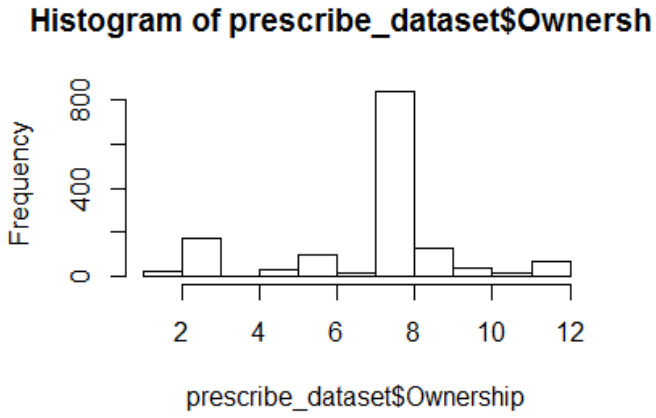
1	Long Island
2	New York City
3	Lower Hudson Valley
4	Capital Region/Northern Catskills
5	Eastern Adirondacks/Lake Champlain
6	Western Adirondacks/Eastern Lake Ontario
7	Central New York
8	Western Finger Lakes
9	Western New York

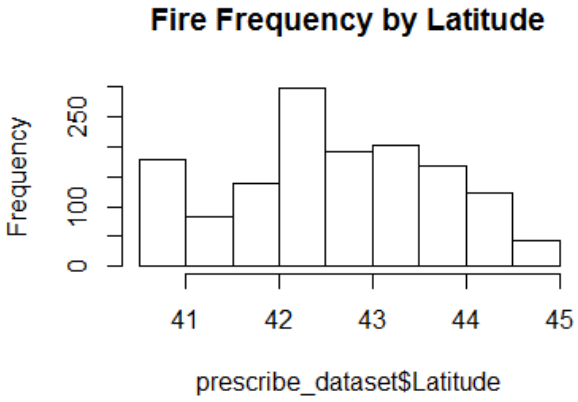


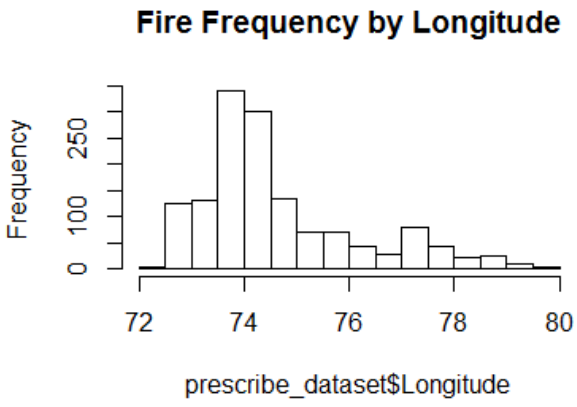


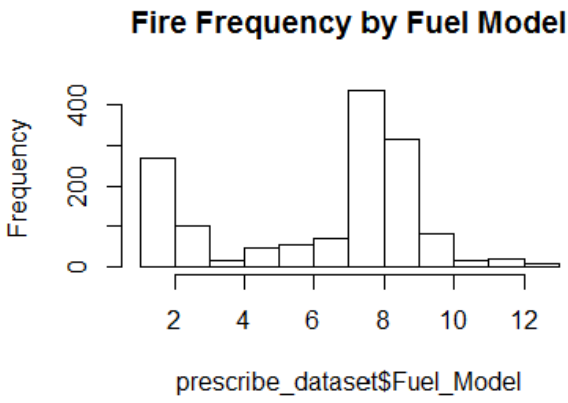
As is clear by the distribution shown, there is no particular decipherable pattern in the data that is revealed. Except that, region 5 shows high frequency of fires and region 3 has big wildfires, we are not able to get much from this information. Also, none of the other histograms show any specific information, or a normal distribution pattern which I was looking for. For ex Ownership distribution by fire incidents.

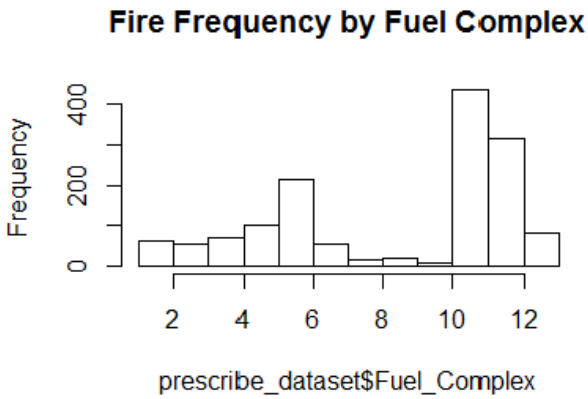
Forest Type Ownership	Factor
Conservation Easement	1
Federal Lands	2
Forest Preserve	3
Native American Lands	4
OPR&HP (State Parks)	5
Other Government	6
Other Not Classified	7
Private Property	8
State Forest/RF Area	9
Unique Area	10
Wilderness Area	11
Wildlife Management Area	12

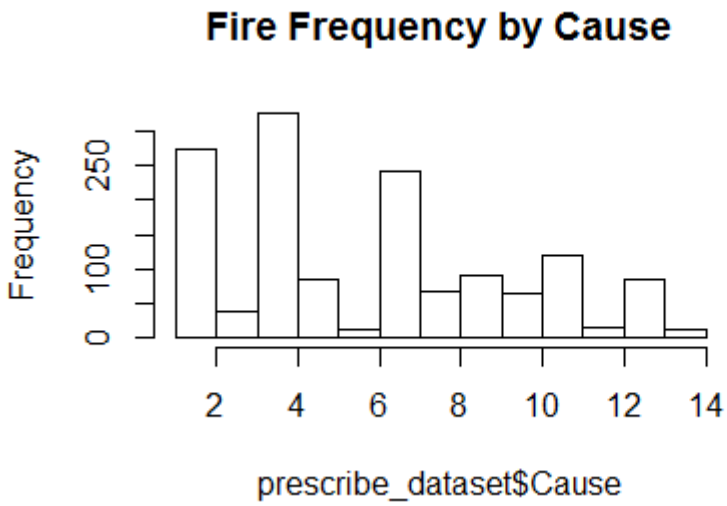






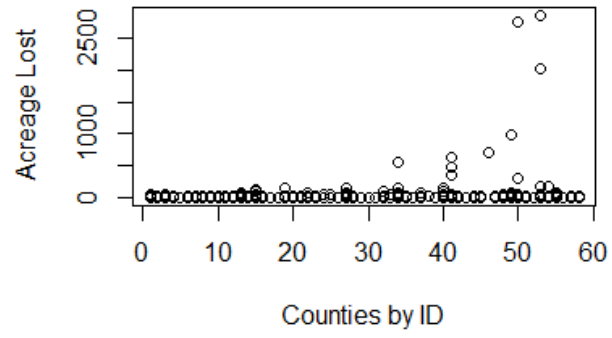






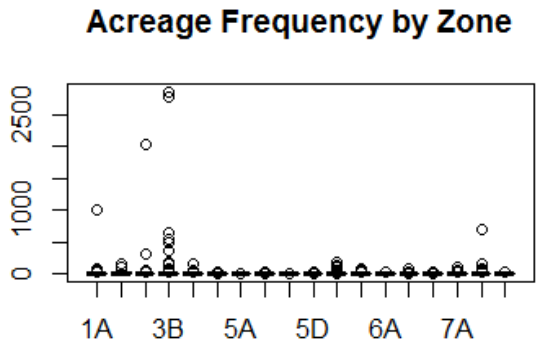
Cause	Cause Numeri
<<Blank>>	1
Campfire	2
Children	3
Debris Burning	4
Equipment	5
Fireworks	6
Incendiary	7
Lightning	8
Miscellaneous	9
Power line	10
Prescribed Fire	11
Railroad	12
Smoking	13
Structure	14

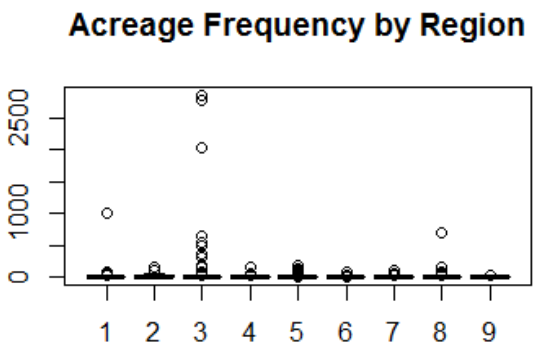
Following this, I was looking to obtain any particular relationship between independent variables, in terms of positive or negative correlation.



The above image is showing the plot between acreage lost versus counties and we can decipher from this image that counties valued between 40-55 have had major wildfire. However these counties are just country names converted to numeric factors, and have no correlation to Zones or regions others

than theirs. For this reason, I decided to **eliminate County column** from the dataset on which I was to perform predictive modelling.





Based on Distribution of fire, it seems clear that except for some big fires favoured in particular counties, the magnitude of fire does not favour any specific county. when seen on a zone basis,higher magnitude fire reduce to one zone, and further reducing to region basis. It seems clear that big fires favour region 3, and then region 5

Statistical Analysis

For further performing statistical analysis in order to gain knowledge about the relation of predictor variables as opposed to response variables, multivariate linear regression was performed. The different references were created for the same datasets, as I proposed to train my models for predicting two different response variables.

The Fire Start Date ,Fire End Date, and Fire Report Date columns were discarded. Instead, three different columns were formed using these columns in order to have a better predicting model. The end date and start date were subtracted from each other to give the **number of days** the fire was on record, minimum being kept at 1 day, for recording incidence. Also the **Fire Start Date** was modified to create a month column and year column, for analytics and training/test set separation by year.

NFFL Fuel Model was also separated into 2 columns, one being Fuel Model, other being Fuel Complex after discussing with Prof. Peter Fox, as a result of which Linear Regression showed an important impact of Fuel Model on the fact whether the fire is prescribed or not.

Further, all the categorical or continuous type variables were converted in the numeric form, such as the **zone, cause, Ownership, Acreage etc.** This was done in order facilitate modelling, as predictive models are usually trained to differentiate datasets based on the overall difference in the vector distances of the predictor variables, just like we find the vector distance between two points in space.

Acreage was split into 5 levels, based on the extent of fire, in order to smooth the prediction interval, as I was considering utilizing supervised models using classification I will elaborate on that further, later.

Result of Multivariate linear Regression for dependent variable *Acreage_Rating* are as **follows** :

```
Call:
lm(formula = fire_svm_dataset$Acreage_Rating ~ ., data = fire_svm_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8930 -0.3053  0.2873  0.5719  1.1531

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.2931539 17.5404622   1.328   0.1844
Region      -0.1198995  0.0637162  -1.882   0.0601 .
Zone         0.0206611  0.0162390   1.272   0.2035
Cause       -0.0439078  0.0068305  -6.428 1.76e-10 ***
Latitude     0.2183907  0.0370323   5.897 4.61e-09 ***
Longitude    0.0285168  0.0463589   0.615   0.5386
Ownership    0.0206320  0.0111849   1.845   0.0653 .
Fuel_Model   0.0145823  0.0075554   1.930   0.0538 .
Total.Days  -0.0004149  0.0005784  -0.717   0.4733
Month        0.0095541  0.0104316   0.916   0.3599
Year        -0.0148494  0.0085405  -1.739   0.0823 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8521 on 1419 degrees of freedom
Multiple R-squared:  0.09091,   Adjusted R-squared:  0.0845
F-statistic: 14.19 on 10 and 1419 DF,  p-value: < 2.2e-16
```

Result of Multivariate Linear Regression for independent variable *Prescribed* are **as follows** :

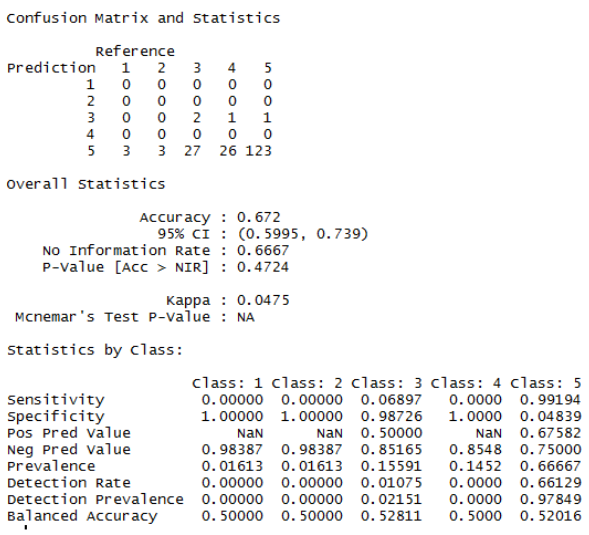
```
Call:
lm(formula = prescribe_dataset$Prescribed ~ ., data = prescribe_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49827 -0.13482 -0.03894  0.06217  0.85813

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.000e+01  4.930e+00   2.028 0.042717 *
Region        6.096e-02  1.789e-02   3.407 0.000675 ***
Zone         -1.441e-02  4.553e-03  -3.164 0.001588 **
Cause         2.922e-02  1.916e-03  15.253  < 2e-16 ***
Latitude     -3.742e-02  1.042e-02  -3.590 0.000341 ***
Ownership     8.420e-03  3.141e-03   2.681 0.007423 **
Fuel_Model   -2.070e-02  2.981e-03  -6.944 5.78e-12 ***
Fuel_Complex -4.142e-03  2.669e-03  -1.552 0.120942
Total.Days    2.555e-05  1.621e-04   0.158 0.874734
Month         1.295e-02  2.929e-03   4.421 1.06e-05 ***
Year         -2.691e-03  2.396e-03  -1.123 0.261555
Acreage       8.399e-06  5.057e-05   0.166 0.868097
Longitude    -4.300e-02  1.309e-02  -3.285 0.001044 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2388 on 1417 degrees of freedom
Multiple R-squared:  0.2652,    Adjusted R-squared:  0.259
F-statistic: 42.63 on 12 and 1417 DF,  p-value: < 2.2e-16
```

For the first image, the multivariate linear regression Model displayed the relation of dependent variable Acreage_ Rating on the rest of the independent variables which remained after all other variables were taken out of consideration. Latitude and Cause were the most significant predictor values as seen by their p-value constants.For the second image, we can see many predictors (p-values) having a strong impact on the response column *Prescribed*, which was created by created by putting prescribed fire *Cause* as 1 and rest as 0.

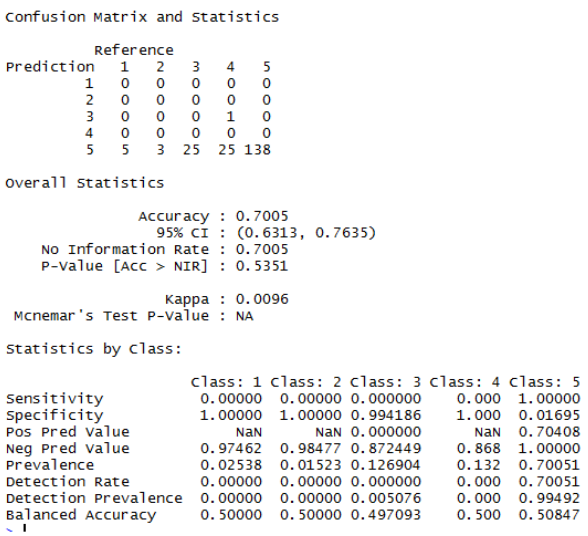
Model Development

For predictive modelling, supervised learning was used as the individual variables were not normally distributed, or showing any correlation to prediction variable. I proceeded first the **Support Vector Machine** which is first preference in this case. This model is used when the data cannot be classified on a linear plane due to the nature of predictor variables.

While predicting the acreage level burnt for 2015, I trained my dataset from 2008-14.Also, while predicting the acreage level burnt for 2016, I trained my dataset from 2008-15. Results as follows:



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1   0   0   0   0   0
         2   0   0   0   0   0
         3   0   0   2   1   1
         4   0   0   0   0   0
         5   3   3  27  26 123

Overall Statistics

               Accuracy : 0.672
                 95% CI : (0.5995, 0.739)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.4724

                  Kappa : 0.0475
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.00000  0.00000  0.06897   0.0000  0.99194
Specificity           1.00000  1.00000  0.98726   1.0000  0.04839
Pos Pred Value            NaN      NaN  0.50000      NaN  0.67582
Neg Pred Value        0.98387  0.98387  0.85165   0.8548  0.75000
Prevalence            0.01613  0.01613  0.15591   0.1452  0.66667
Detection Rate        0.00000  0.00000  0.01075   0.0000  0.66129
Detection Prevalence  0.00000  0.00000  0.02151   0.0000  0.97849
Balanced Accuracy     0.50000  0.50000  0.52811   0.5000  0.52016
```



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1   0   0   0   0   0
         2   0   0   0   0   0
         3   0   0   0   1   0
         4   0   0   0   0   0
         5   5   3  25  25 138

Overall Statistics

               Accuracy : 0.7005
                 95% CI : (0.6313, 0.7635)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 0.5351

                  Kappa : 0.0096
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.00000  0.00000 0.000000     0.000  1.00000
Specificity           1.00000  1.00000 0.994186     1.000  0.01695
Pos Pred Value            NaN      NaN 0.000000       NaN  0.70408
Neg Pred Value        0.97462  0.98477 0.872449     0.868  1.00000
Prevalence            0.02538  0.01523 0.126904     0.132  0.70051
Detection Rate        0.00000  0.00000 0.000000     0.000  0.70051
Detection Prevalence  0.00000  0.00000 0.005076     0.000  0.99492
Balanced Accuracy     0.50000  0.50000 0.497093     0.500  0.50847
```

With Support Vector Modelling above, I did not get significant accuracy. I tried many other variations like reducing the number of levels based on their distribution, but it was not very helpful.

Following this, I predicted which fires were prescribed in 2015, by training my dataset from 2008-14. I also predicted which fires were prescribed in 2016, by training my dataset from 2008-15.

The prediction was a binary outcome which resulted in good overall prediction accuracy against the test set. However, after having a look at confusion matrix plot, it was visible that the accuracy of prediction for prescribed fire was very poor. The *false negative* zone of the matrix, which showed that the actual outcome versus predicted outcome was zero figure (for 2014-15), and 2 outcomes (for 2015-16) . **Results below:**

```
> confusionMatrix(predict_set, prescribe_test_S
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 173  11
         1   2   0

               Accuracy : 0.9301
                 95% CI : (0.8834, 0.9623)
    No Information Rate : 0.9409
    P-Value [Acc > NIR] : 0.7870

                  Kappa : -0.0185
 Mcnemar's Test P-Value : 0.0265

            Sensitivity : 0.9886
            Specificity : 0.0000
         Pos Pred Value : 0.9402
         Neg Pred Value : 0.0000
             Prevalence : 0.9409
         Detection Rate : 0.9301
   Detection Prevalence : 0.9892
      Balanced Accuracy : 0.4943

       'Positive' Class : 0
```

```
> confusionMatrix(predict_set, prescribe_test_S
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 184  11
         1   0   2

               Accuracy : 0.9442
                 95% CI : (0.9023, 0.971
    No Information Rate : 0.934
    P-Value [Acc > NIR] : 0.346194

                  Kappa : 0.2535
 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 1.0000
            Specificity : 0.1538
         Pos Pred Value : 0.9436
         Neg Pred Value : 1.0000
             Prevalence : 0.9340
         Detection Rate : 0.9340
   Detection Prevalence : 0.9898
      Balanced Accuracy : 0.5769

       'Positive' Class : 0
```

Following this, I decided to train a prediction model using **Random Forest Classification**. I chose random forest Classification for a few reasons.

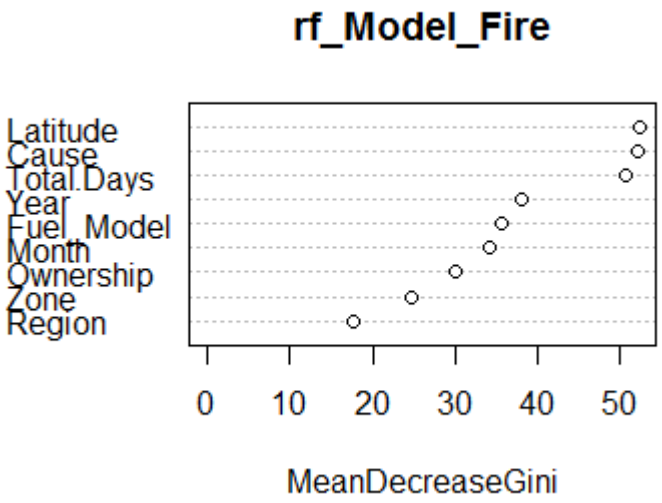
- With only two impacting variables for acreage levels, it was difficult to find the right kernel, which would give proper predictions
- Random Forest gives a more robust prediction, because it's a tree based classifier which grows in complexity.

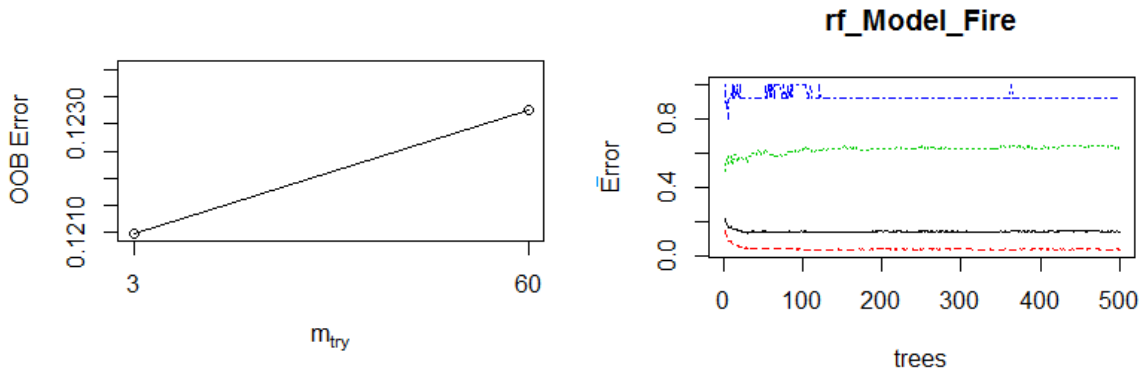
Following results were obtained while predicting acreage level burnt for year 2015/2016 against actual data.

```
> confusionMatrix(predict_set, fire_test_Set$Acreage_Rating)
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1   0   0   0   0   0
         2   0   0   1   1   0
         3   0   0  10   4   4
         4   1   0   3   2   0
         5   2   3  15  20 120

Overall Statistics

               Accuracy : 0.7097
                 95% CI : (0.6388, 0.7738)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.121

                  Kappa : 0.2859
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.00000  0.00000  0.34483  0.07407   0.9677
Specificity           1.00000  0.98907  0.94904  0.97484   0.3548
Pos Pred Value            NaN  0.00000  0.55556  0.33333   0.7500
Neg Pred Value        0.98387  0.98370  0.88690  0.86111   0.8462
Prevalence            0.01613  0.01613  0.15591  0.14516   0.6667
Detection Rate        0.00000  0.00000  0.05376  0.01075   0.6452
Detection Prevalence  0.00000  0.01075  0.09677  0.03226   0.8602
Balanced Accuracy     0.50000  0.49454  0.64694  0.52446   0.6613
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1   0   0   0   0   0
         2   0   0   0   0   0
         3   4   1   6   7   6
         4   0   0   5   1   7
         5   1   2  14  18 125

Overall Statistics

               Accuracy : 0.6701
                 95% CI : (0.5997, 0.7352)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 0.844

                  Kappa : 0.1891
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.00000  0.00000  0.24000 0.038462   0.9058
Specificity           1.00000  1.00000  0.89535 0.929825   0.4068
Pos Pred Value            NaN      NaN  0.25000 0.076923   0.7812
Neg Pred Value        0.97462  0.98477  0.89017 0.864130   0.6486
Prevalence            0.02538  0.01523  0.12690 0.131980   0.7005
Detection Rate        0.00000  0.00000  0.03046 0.005076   0.6345
Detection Prevalence  0.00000  0.00000  0.12183 0.065990   0.8122
Balanced Accuracy     0.50000  0.50000  0.56767 0.484143   0.6563
```

Compared to Support Vector Machine, there is no marked difference in overall accuracy, but confusion plot matrix is much better in this case, and the model is able to predict true positives in this case. Following is the varImpPlot for Acreage Rating which is mostly similar for prediction for both years : 2015 & 2016. RF model plot describing error with reference to trees is also given.





I tried **tuneRF function**, which gave me mtry value of 3 for lowest error, but it did not increase the highest value of accuracy for both prediction models when the formula was modified with mtry.

The varimpPlot gives us the significant variables which a are used to make a prediction about acreage Level, latitude being the most determinant of all, followed by Cause and Total Days. Although the model was not able to predict category 1 and category 2 fires accurately, it worked comparatively better in predicting level 3 and level 4 fires as opposed to SVM, which could only predict category level 5 fires, but did not predict other categories.

Random Forest worked very effectively in predicting if the fire was prescribed or not. **Results below**

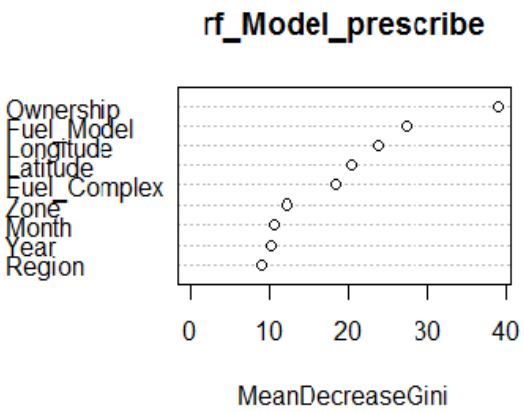
```
> confusionMatrix(predict_set, prescribe_tes
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 169   0
         1   6  11

               Accuracy : 0.9677
                 95% CI : (0.9311, 0.9881)
    No Information Rate : 0.9409
    P-Value [Acc > NIR] : 0.07250

                  Kappa : 0.7691
 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.9657
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6471
             Prevalence : 0.9409
         Detection Rate : 0.9086
   Detection Prevalence : 0.9086
      Balanced Accuracy : 0.9829

       'Positive' Class : 0

>
```

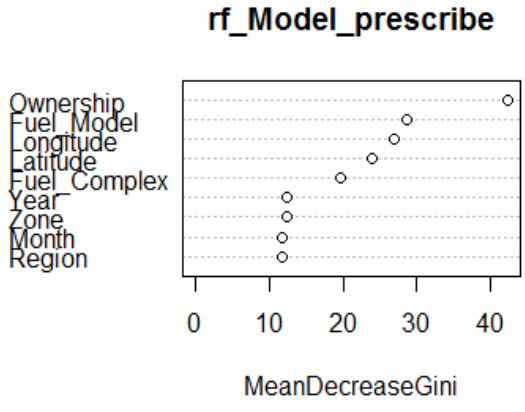
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 183   5
         1   1   8

               Accuracy : 0.9695
                 95% CI : (0.9349, 0.9887)
    No Information Rate : 0.934
    P-Value [Acc > NIR] : 0.02245

                  Kappa : 0.7117
 Mcnemar's Test P-Value : 0.22067

            Sensitivity : 0.9946
            Specificity : 0.6154
         Pos Pred Value : 0.9734
         Neg Pred Value : 0.8889
             Prevalence : 0.9340
         Detection Rate : 0.9289
   Detection Prevalence : 0.9543
      Balanced Accuracy : 0.8050

       'Positive' Class : 0

>
```

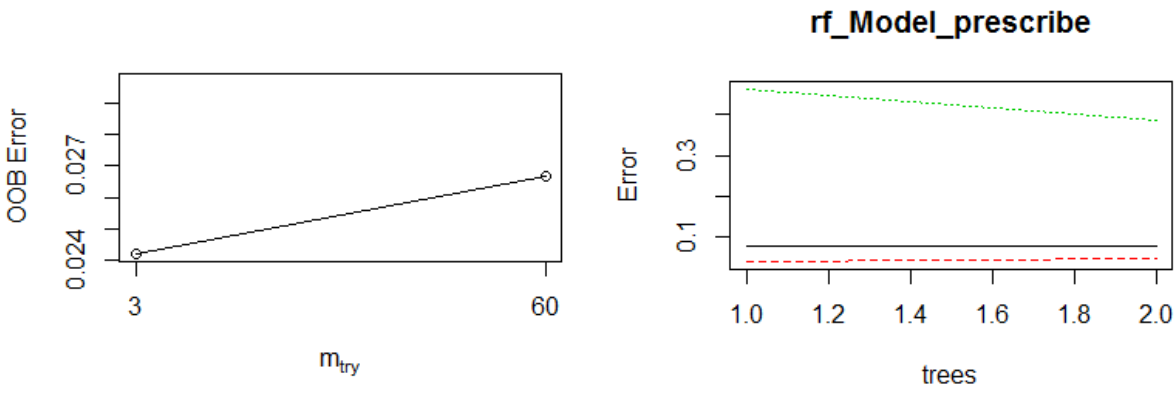
The overall accuracy for predicting if fire was prescribed or not, for year 2015 (left) and year 2016 (right) was very much spot on. Total Days of fire, and acreage / acreage_rating had no significant impact on prescribed variable as observed by their high p-value in linear regression plot. They were functioning as noise in the linear regression.

We can also notice that the balanced accuracy is very high when predicting for year 2015. Here we can say that the choice of predictor variables, combined with Random Forest classification, was a success overall, and this has been the best prediction from this dataset so far.

The **varImpplot plot below** describes the significance of variable in constructing the random forest tree for classifying if the fire was *prescribed* or not. Ownership, Fuel_Model, and geographic parameters clearly take precedence over other parameters.







I tried **tuneRF function**, which gave me mtry value of 3 for lowest error, but it did not increase the highest value of accuracy for both prediction models when the formula was modified with mtry

The last classification model I trained my dataset with, was Naive Bayes classifier, which can be used when you are training with less number of data in your training set. As I had 1430 rows in total in my dataset, I supposed it would make sense to apply Naive Bayes Classification. **Following** are the results I obtained for Acreage Rating.

```
> confusionMatrix(predict_NB, fire_test_Set$Acreage_Rating)
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1  1  0  1  0  1
         2  0  0  0  0  1
         3  0  2 23 18 80
         4  0  0  3  0  1
         5  2  1  2  9 41

Overall Statistics

               Accuracy : 0.3495
                 95% CI : (0.2812, 0.4227)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0657
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity          0.333333 0.000000   0.7931  0.00000   0.3306
Specificity          0.989071 0.994536   0.3631  0.97484   0.7742
Pos Pred Value       0.333333 0.000000   0.1870  0.00000   0.7455
Neg Pred Value       0.989071 0.983784   0.9048  0.85165   0.3664
Prevalence           0.016129 0.016129   0.1559  0.14516   0.6667
Detection Rate       0.005376 0.000000   0.1237  0.00000   0.2204
Detection Prevalence 0.016129 0.005376   0.6613  0.02151   0.2957
Balanced Accuracy    0.661202 0.497268   0.5781  0.48742   0.5524
```

```
> confusionMatrix(predict_NB, fire_test_Set$Acreage_Rating)
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1  2  1  0  1  1
         2  0  0  0  0  0
         3  2  1 17 13 53
         4  0  1  2  4 23
         5  1  0  6  8 61

Overall Statistics

               Accuracy : 0.4264
                 95% CI : (0.3564, 0.4987)
    No Information Rate : 0.7005
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1224
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.40000  0.00000  0.68000   0.1538   0.4420
Specificity           0.98438  1.00000  0.59884   0.8480   0.7458
Pos Pred Value        0.40000      NaN  0.19767   0.1333   0.8026
Neg Pred Value        0.98438  0.98477  0.92793   0.8683   0.3636
Prevalence            0.02538  0.01523  0.12690   0.1320   0.7005
Detection Rate        0.01015  0.00000  0.08629   0.0203   0.3096
Detection Prevalence  0.02538  0.00000  0.43655   0.1523   0.3858
Balanced Accuracy     0.69219  0.50000  0.63942   0.5009   0.5939
```

The prediction accuracy is very low as compared to SVM or Random Forest, however, the model makes sincere efforts in predicting through all the categorical variables, even predicting the category 1 fires, which the rest two modes could not predict.

It is possible that with better or lesser choice of variables, the result could have been improved. **Naive Bayes** is commonly used for text classification, because it cannot represent complex behavior as opposed to Random Forest. This inability is seen in terms of its overall inaccuracy in predicting Acreage Rating. **Below**, we have for *Prescribed*

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 176  10
         1   8   3

               Accuracy : 0.9086
                 95% CI : (0.8594, 0.9449)
    No Information Rate : 0.934
    P-Value [Acc > NIR] : 0.9369

                  Kappa : 0.2017
 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.9565
            Specificity : 0.2308
         Pos Pred Value : 0.9462
         Neg Pred Value : 0.2727
             Prevalence : 0.9340
         Detection Rate : 0.8934
   Detection Prevalence : 0.9442
      Balanced Accuracy : 0.5936

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 156   9
         1  19   2

               Accuracy : 0.8495
                 95% CI : (0.7898, 0.8976)
    No Information Rate : 0.9409
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.0514
 Mcnemar's Test P-Value : 0.08897

            Sensitivity : 0.89143
            Specificity : 0.18182
         Pos Pred Value : 0.94545
         Neg Pred Value : 0.09524
             Prevalence : 0.94086
         Detection Rate : 0.83871
   Detection Prevalence : 0.88710
      Balanced Accuracy : 0.53662

       'Positive' Class : 0
```

Summary of **Naive Bayes** for predicting the likelihood of fire being prescribed in 2016 is **as follows**:

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.91321979 0.08678021

Conditional probabilities:
   Region
Y       [,1]     [,2]
  0 4.918295 2.186114
  1 3.887850 2.611033

   Zone
Y       [,1]     [,2]
  0 9.448490 5.600358
  1 6.747664 6.231529

   Latitude
Y       [,1]      [,2]
  0 42.64982 1.1002848
  1 42.05421 0.9456235

   Ownership
Y       [,1]     [,2]
  0 7.402309 2.053670
  1 8.719626 2.550599

   Fuel_Model
Y       [,1]     [,2]
  0 6.792185 3.070052
  1 3.102804 2.370951
```

```
   Fuel_Model
Y       [,1]     [,2]
  0 6.792185 3.070052
  1 3.102804 2.370951

   Fuel_Complex
Y       [,1]     [,2]
  0 9.159858 3.389294
  1 5.747664 2.355617

   Month
Y       [,1]     [,2]
  0 5.460036 2.280293
  1 5.168224 2.296371

   Longitude
Y       [,1]     [,2]
  0 74.75311 1.550015
  1 74.32804 1.704865

   Year
Y       [,1]     [,2]
  0 2011.515 2.413157
  1 2011.299 2.427107

> |
```

For the likelihood of fire being prescribed or not the accuracy was low as compared to Random Forest or SVM. In case of predicting binary outcomes, it seems that Random Forest was the best approach, as it generates a classification tree based on each vector. The left side represents prediction for year 2015, and the right side represents prediction for year 2016.

Summary of Naive Bayes for predicting Acreage Rating in 2016 is **as follows**:

```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        1         2         3
0.83049473 0.15896188 0.01054339

Conditional probabilities:
   Region
Y       [,1]     [,2]
  1 4.855469 2.165356
  2 4.750000 2.622144
  3 3.923077 2.100061

   Zone
Y       [,1]     [,2]
  1 9.291992 5.533661
  2 8.974490 6.536239
  3 6.692308 5.452593

   Cause
Y       [,1]     [,2]
  1 6.049805 3.324494
  2 7.887755 3.474102
  3 6.846154 4.336784

   Latitude
Y       [,1]      [,2]
  1 42.69355 1.1046146
  2 42.14082 0.9443855
  3 41.97692 1.0771520

   Ownership
Y       [,1]     [,2]
  1 7.439453 2.087585
  2 8.020408 2.202468
  3 6.000000 3.082207

   Fuel_Model
Y       [,1]     [,2]
  1 6.663086 3.113802
  2 5.387755 3.389676
  3 7.769231 2.586949

   Total.Days
Y       [,1]       [,2]
  1 3.815430 46.119867
  2 3.193878  2.912953
  3 9.692308  4.230536

   Month
Y       [,1]     [,2]
  1 5.496094 2.248489
  2 5.045918 2.364900
  3 6.461538 2.989297

   Year
Y       [,1]     [,2]
  1 2011.454 2.405972
  2 2011.679 2.458632
  3 2012.077 2.361551

> |
```

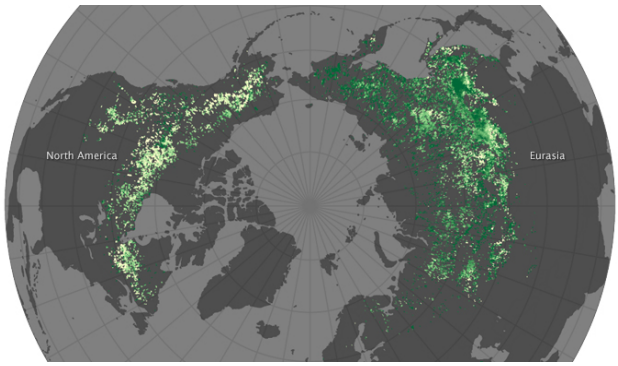
Following is the Acreage rating which I formulated, based on the dataset. Below that is the cause of fire (categorical variable)

Fire Category	Limits
1	Acreage > 100 Acres
2	Acreage between 75 to 100 Acres
3	Acreage between 50 to 75 Acres
4	Acreage between 25 to 50 Acres
5	Acreage between 0 to 25 Acres

Category	Cause
0	Not prescribed. Other causes such as debris burning, smoking, campfire etc
1	Prescribed Fire (Started by Forest Rangers)

Summary and Conclusion

One interesting observation made during Data Analysis was that increasing Latitude was somewhat positively correlated to Acreage Burnt, which was unusual considering that we're going further north. Even if caused by human error, the acreage burnt should still not be high. This variation has been observed by NASA as well, and they attribute it to global increase in CO2 levels for the way high altitude forest are behaving in North America and Eurasia. Citations are present in references.



Based on the overall modelling, there were many conclusions that I could draw:

1. Firstly, supervised learning is not very efficient when the values being predicted are not binary, and very categorical. I could have explored neural networks as well, but working with first three supervised models led to a conclusion that supervised models were not highly efficient, at least when applied to make dataset. Unsupervised learning, like K-means clustering could have been explored.

2. The size of the dataset was very small, due to which Random Forest was not very efficient in predicting Acreage Rating. With a large dataset, there would have been possible to better classify Acreage Rating into specific subsets. The classifier worked well for prediction of binary outcome *Prescribed*

In summary, we can not prefer any classification method if it outperforms others in one context and fails in other context. There could be some patterns that can be extracted from data, or they could be patterns we are trying to find that just don't exist. In my datasets, I was able to fortunate to observe both possibilities. The overall accuracy result is as follows :

	SVM	Random Forest	Naive Bayes
Prescribed Fire (2015)	93.01%	96.77%	84.95%
Prescribed Fire (2016)	94.42%	96.95%	90.86%
Acreage Burnt (2015)	67.2%	70.97%	34.95%
Acreage Burnt (2016)	70.05%	67.01%	42.64%

References

1. New York State Forest Ranger WildLand Fire Reporting https://data.ny.gov/Energy-Environment/New-York-State-Forest-Ranger-Wildland-Fire-Reporti/miub-n5th/data
2. Metadata Information for Dataset https://data.ny.gov/Energy-Environment/New-York-State-Forest-Ranger-Wildland-Fire-Reporti/miub-n5th/about
3. High-Latitude Forest Fires Behave Differently in North America and Eurasia https://earthobservatory.nasa.gov/IOTD//view.php?id=85172
4. WildFires: A Symptom of Climate Change https://www.nasa.gov/topics/earth/features/wildfires.html
5. WildFires: NYS Dept of Environmental Conservation http://www.dec.ny.gov/lands/4975.html
6. New York State Forest Ranger Wildland Fire Reporting Incident Map https://data.ny.gov/w/b7g8-5ywk/caer-yrtv?cur=A5qbR0CDOJr&from=root