In this project, I analyzed the forest fire dataset from UCI learning repository using linear regression, random forests and gbm. The reason why I carried out this project is I wanted to find out the best machine learning method to use on this dataset.

The dataset consisted of 13 variables and 517 observations with the response variable being area, the amount that the forest burned in ha. The 12 explanatory variables are: X, Y, month, day, FFMC, DMC, DC, ISI, temp, RH, wind, and rain.

I began the project by cleaning the data so that it was suitable for use. I first started by removing any NA values within the dataframe. From there, I performed a logarithmic transformation on the column area because the data was skewed towards zero. In order to prevent errors in the logarithmic transformation, I added one to each observation so that there would not be any logarithmic transformation of the number zero which leads to a computational error as log10(0) does not exist. I finally read in the categorical variables month, and day as a numeric factor and now the dataset is ready for use.

I created my own kfolds splitting function and then utilized it so that I could compare the different models efficiently. I compared the RMSE of four different types of models: full model linear regression, best subset (stepwise) regression, random forest, and gbm. The best subset used X, DMC, RH, and wind to predict the response variable. For the random forest method, and the gbm method, I started off by using the default parameters so that there wasn't any bias in the results. The results are as follows:

	1	2	3	4	5	Mean
linear_mod	0.4190598	0.5252311	0.3243039	0.2643307	0.4749551	0.4015761
best_subset	0.4245284	0.5162507	0.3334546	0.2706593	0.4900462	0.4069878
rf	0.4223552	0.5416939	0.3111166	0.2852410	0.4709003	0.4062614
gbm	0.4335821	0.5017957	0.2967280	0.2643825	0.4887196	0.3970416

From the results, we can see that gbm performed the best with the lowest RMSE after taking the average of the iterations.

From there, for the random forests and stochastic gradient boosting methods, I tuned the parameters so that I knew which parameters provided the best results. The resulting best models are as follows:

```
Random Forest

363 samples
 12 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 363, 363, 363, 363, 363, 363, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared    MAE
   2    0.6391249  0.01728883  0.5153656
   7    0.6513197  0.01697599  0.5220249
  12    0.6555992  0.01642680  0.5229557

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
```

```
Stochastic Gradient Boosting

363 samples
 12 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 363, 363, 363, 363, 363, 363, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE       Rsquared     MAE
  1                   50      0.6410369  0.008837779  0.5244630
  1                  100      0.6501560  0.008664605  0.5264929
  1                  150      0.6573393  0.007981437  0.5297877
  2                   50      0.6470630  0.011517580  0.5232538
  2                  100      0.6648695  0.009223227  0.5304320
  2                  150      0.6779643  0.008242935  0.5380617
  3                   50      0.6589468  0.007952060  0.5294060
  3                  100      0.6785864  0.007203440  0.5398160
  3                  150      0.6922043  0.006686523  0.5463451

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 50, interaction.depth = 1, shrinkage = 0.1 and n.minobsinnode = 10.
```

I then tested the gbm method and the random forest methods with the optimal values for their adjustable parameters. I once again compared the RMSE of the newly improved models and the results are now as follows.

```
                1         2         3         4         5      Mean
rf_tuned  0.3916185 0.3922957 0.3925345 0.3862511 0.3952985 0.3915996
gbm_tuned 0.3493168 0.3288531 0.3377799 0.3349117 0.3423126 0.3386348
```

Overall, the methods: linear regression, best_subset, random forest all had similar RMSE values all around 0.40. Gbm with the default parameters and random forests with the tuned parameters had a slightly better RMSE of 0.39. The gbm method with tuned parameters had the best results with an average RMSE of 0.3386348. There are many more machine learning techniques, but in this project, of the six tested models, the gbm method with tuned parameters had by far the best fit.