Model Building with Forest Fire Data: Data Mining, Exploratory Analysis and Subset Selection

By Hye Rin Kim

Supervised by Dr. McLeod

Master's Project

August 2009

Contents

1 Introduction 1

2 Description of Forest Fire Data 2

3 Standardization and Error Criteria 4

4 Analysis 5

4.1 Exploratory Analysis 5
4.2 Multiple Linear Regression 12
4.3 Neural Networks 14
4.4 Support Vector Machine 16
4.5 Random Forest 18
4.6 LASSO and LAR 19
4.7 Summary of Results 22
4.8 Regression Error Characteristic (REC) Curve 23
4.9 Other Attempts: Logistic Regression 25
4.9.1 Logistic Regression 25
4.9.2 Multi response Logistic Regression 28

5 Conclusion 30

Abstract

Forest fires are a major environmental issue, creating economical and ecological damage while dangering human lives. Cortez and Morais (2007) provided data on 517 forest fires in northeast region of Portugal. Cortez and Morais (2007) claimed fitted a Support Vector Machine which they claimed outperformed Multiple Linear Regression for prediction. In particular, Cortez and Morais (2007) claimed to be able to predict small forest fires based on a Regression Error Characteristic (REC) Curve. A number of input variables are available including spatial, temporal and weather attributes. Cortez and Morais (2007) also explored the use of other data mining methods including Feed-Forward Neural Nets and Random Forests. A brief review of these data mining methods is given. It is of interest to determine which of the inputs are most relevant as well as to assess the quality of any predictions that can be made. We find a number of shortcomings in the formulation and analysis by Cortez and Morais (2007). An improved formulation of the problem is suggested. Logistic and multi-response logistic regression are also suggested. We will look at Cortez and Morais paper critically. We perform a Data Mining (DM) approach to predict the burned area of forest fires. Eight different DM techniques such as Naive, Multiple Regression, Feed Forward Neural Networks, a skip-layer Neural Networks, Support Vector Machine, LASSO and Lar, and Random Forest and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes) will be applied on recent real-world data collected from the northeast region of Portugal. We will also take a different approach as a classification problem to predict the burned area of forest fires. Logistic Regression and Multiclass logistic regression will be applied to see if they provide improvements on predicting forest fires data and which input variables are important.

Chapter 1

Introduction

In section 2, the description of the forest fires data will be presented. Exploring data will be present in section 4.1 and some modifications will be provided on some data to make improvements and explore which variables are important.

In section 4, all data mining methods and analyses will be presented. Data Mining methods such as Naive, Multiple Regression, Feed Forward Neural Networks, a skip-layer Neural Networks, Support Vector Machine, LASSO and Lar, and Random Forest and eight distinct feature selection setups are applied on forest fires data. In Cortez and Morais paper [3], they used 4 features selection setups but we added 4 more feature selection setups. Some modifications on spatial and temporal variables are provided. We will compare our results with the paper. Later, we will compare the results with MAD and RMSE as well as REC curve and figure out which method performs the best for predicting forest fires and which input variables are important. Also, Logisistic and Multi-response Logistic Regression will be applied as a classification problem and see if they make some improvements.

Chapter 2

Description of Forest Fire Data

This problem will consider forest fire data from the Montesinho natural park, from the Trá-os-Montes Northeast region of Portugal. Satellite-based and infrared/smoke scanners have high costs. However, weather conditions, such as temperature and air humidity, are known to affect fire occurence, automatic meteorological satations are often available, and such data can be collected in real-time with low costs. We will present a Data Mining forest fire approach with emphasis on the use of real-time and non-costly meteorological data to predict the burned area of forest fires.

The data was collected from January 2000 to December 2003 with a total of 517 entries. In the dataset, there are 13 attributes that are the spatial and temporal attributes, four FWI components that are affected directly by the weather conditions, four meteorological attributes, and the response variable, the burned **area**. The data is consist of 12 input variables that are X, Y, month, day, FFMC, DMC, DC, ISI, temp, RH, wind, and rain and the response variable that is area.

The first four attributes are the spatial and temporal attributes. The first two attributes are the X and Y axis values where the fire occured within a 9*9 grid and the third and fourth attributes are the month and day of the week temporal variables. The next four FWI components are affected directly by the weather conditions. The forest Fire Weather Index(FWI) [3] is the Canadian System for rating fire danger and it includes six components. Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Different scales are used for each of the FWI elements, but high values suggest more severe burning conditions. The BUI and FWI were discarded since they are dependent of the previous values. The next four weather attributes are used by the FWI system and from the meteorological

Attributes	**Description**
X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
month	Month of the year (January to December)
day	Day of the week (Monday to Sunday)
FFMC	FFMC code
DMC	DMC code
DC	DC code
ISI	ISI index
temp	Outside temperature (in Celsius)
RH	Outside relative humidity (in percentage)
wind	Outside wind speed (in kilometer per hour)
rain	Outside rain (in millimeter per square meter)
area	Total burned area (in ha)

station database. In this case the values denote instant records, as given by the station sensors when the fire was detected. The rain variable represents the accumulated precipitation within the previous 30 minutes. The area variable represents the total burned area in hectares (ha). In the dataset, there are 247 samples with a zero value. All entries denote fire occurrences and zero value means that an area lower than $1ha/100 = 100m^2$ was burned. The burned area denoted a positive skew and we applied the logarithm transformation, y = ln(x + 1), to reduce the skewness and improve symmetry.

We'd like to examine the impact of the input variables and four distinct feature selection setups were tested for each DM algorithm. The four feature selection setups are as follows.

```
STFWI  using spatial, temporal and the four FWI components
STM    with the spatial, temporal and four weather variables
FWI    using only the four FWI components
M      with the four weather conditions
```

Later, we will make some modificaitons on spatial and temporal variables and examine four more feature selection setups such as:

```
SF   using spatial and the four FWI components
SM   using spatial and four weather variables
SFM  using spatial, the four FWI compoents, and four weather variables
S    using spatial variables only
```

Chapter 3

Standardization and Error Criteria

Before performing analyses and fitting the models, some preprocessing was required. All attributes except the response variable were standardized to a zero mean and one standard deviation. After fitting the models, the overall performace is computed using the Mean Absolute Deviation (MAD) and Root Mean Squared (RMSE) below.

$$MAD = 1/N * \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N}$$

Chapter 4

Analysis

In this chapter, first we will perform some exploratory analysis on input variables and make some improvments on some input variables. Then, the Data Mining methods and their analyses will be provided. We will predict the forest fires using all Data Mining methods, provide the overall performace by using MAD and RMSE criteria and the REC curves, and later compare all the DM methods.

4.1 Exploratory Analysis

We will perform exploratory analysis on input attributes for forest fires data.

We perform explanatory analysis to see how important the input variables are. We solve this using regression. We use F test using anova function in R. First, we analyze spatial variables of X and Y and see if they are important for predicting 'lburned'. We standardized X and Y coordinates and fitted a F test.

From Table 4.1, X and Y are not significant and the R squared is 0.39%. The linear predictors X and Y used by Cortez and Morais are not useful and the result is shown in Table 4.2.

To make an improvements on the spaital variables of X and Y, we tried and performed the F test for 'xyarea' factor variable for the spatial variables. The 'xyarea" variable is a factor variable which indicates each area on the grid. The result of the F test on the 'xyarea' factor variable is shoWn in table 4.2. The factor variables are statistically significant at 1.13% We also tried F test using interaction and the interaction was not important. So an additive model works.

Also, we tried to use additive regression splines to make an improvement on the spatial variables. We suggest to use additive regression splines on spatial variables X and Y. We make 7 splines for each spatial variables X and Y and test the splines for each spatial variables X and Y. The result is shown in Table 4.2. We make additive spline basis using ns function in splines package in R. We standardized all splines and fitted the F test. The last column which is Y7 is redundant, so it is removed and refitted the F test. The result is shown in Table 4.3.

Using splines for spatial variable is highly significant at 0.7% in Table 4.3 The R squared is 6.73%, it has a little predictive power. The factors have a higher R squared which is 10.79% than the splines but they have a lower p-value. Using the factors don't make a real difference with using the splines. So, we suggest to use additive regression splines.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesXY	2	3.92	1.96	1.00	0.3678
Residuals	514	1005.18	1.96		

Table 4.1: Results of X and Y coordinates

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesFactor	35	108.93	3.11	1.66	0.0113
Residuals	481	900.17	1.87		

Table 4.2: Results of X and Y factor variables

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesS	13	67.86	5.22	2.79	0.0007
Residuals	503	941.24	1.87		

Table 4.3: Results of X and Y splines

For more datails, we would like to visualize if there is a regional variation in the fires and see how best the regional variation can be modelled. We want to explore if the number of fires are dependent on X and Y.

In figure 4.1, X and Y coordinates against the number of fires are plotted and the number of fires are not dependent on X and Y. RGB color is encoded for quartile information. Black is (lowest) for observations less than or equal to the first quartile. Red (lowest frequency), Green and Blue (highest frequency) are corresponding the next quartiles.

We also performed F test on X and Y factors vs. the number of fires and the result is in table 4.4. They are not significant at all.

Also, X and Y against the median of the 'lburned' for each region are plotted in Figure 4.2. Median 'lburned' depends on X and Y. That means that low X, Y tends to have low 'lburned'. So, there is obvious dependence of X and Y on area for 'lburned'. We also performed the F test on X and Y factors vs. the the median of the 'lburned' and the result is shown in table 4.5. They are statistically significant now.

Also, X and Y is plotted against the mean of 'lburned' for each region in Figure 4.3. Mean 'lburned' depends on X and Y. That means that low X, Y tends to have low 'lburned'. We also performed the F test on X and Y factors vs. the mean of the 'lburned' and the result is shown in table 4.5. They are statistically significant now.

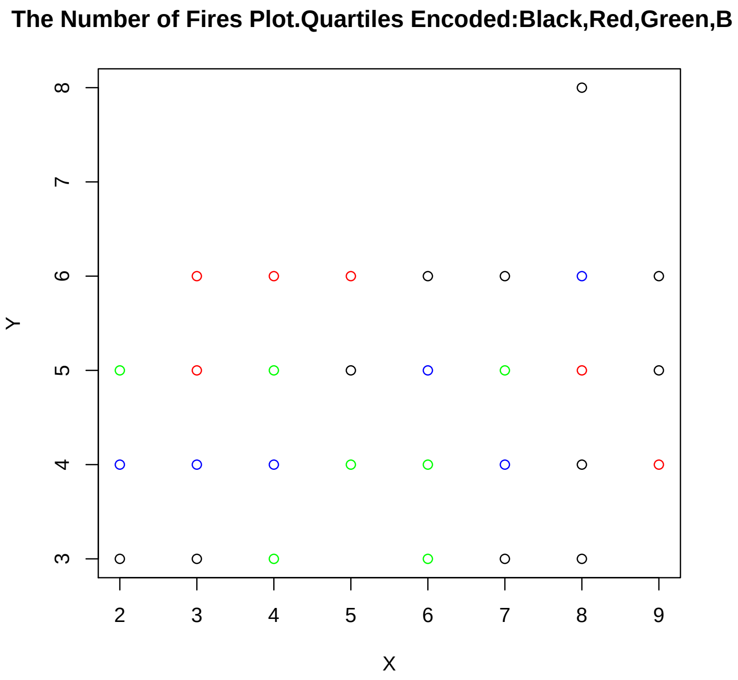

Figure 4.1: The Number of Fires Plot

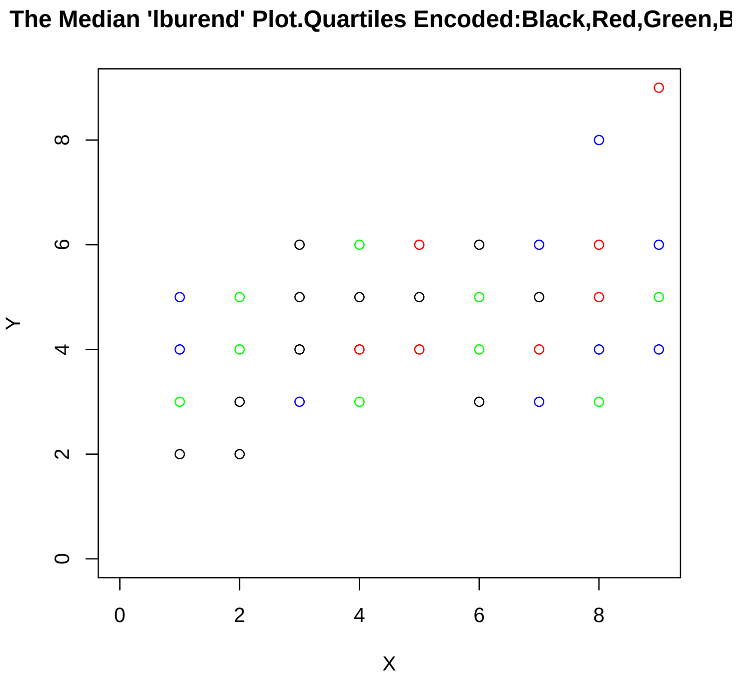

Figure 4.2: The Median lburned Plot

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	8	1145.51	143.19	0.55	0.8049
Y	6	1427.30	237.88	0.92	0.5031
Residuals	21	5455.50	259.79		

Table 4.4: The result of the number of fires.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	8	13.25	1.66	2.30	0.0603
Y	6	22.15	3.69	5.13	0.0022
Residuals	21	15.12	0.72		

Table 4.5: The result of median lburned.

We also plotted X and Y against the SD of 'lburned' for each region and SD 'lburned' depends on X but not on Y. Thus, there is an obvious dependence of X and Y on the response variable 'lburned' and there is a regional variation in the fires.

However, the X and Y coordinates are not significant as it is shown in table 4.1. So, we suggest to use tensor splines rather than using simple linearlity of X and Y values. We tried and performed F test for the 7 splines of the spatial variables earlier in the previous table 4.3 and they are statistically significant at 0.07%. Therefore, the 7 splines for spatial variables X and Y are important for impacting the 'lburned' of the forest fires. So, the regional variation can be modelled by the 7 splines of X and Y.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	8	10.62	1.33	2.33	0.0577
Y	6	17.02	2.84	4.98	0.0026
Residuals	21	11.97	0.57		

Table 4.6: The result of mean lburned.

Next, we perform exploratory analysis for temporal variables. We first analyze the temporal variables of month and day. We performed the F test using month and day categorical variable in in Table 4.9 and the temporal variables are not significant. So, we tried to use sinusoids for temporal variables of month and day. We performed F test for twi sinusoids for each temporal variables using anova and the result is shown in Table 4.7. They are not statistically significant. In Table 4.8, we used only month sinusoids and performed the F test. Using month sinusoids only provided better p-value than using both month and day sinusoids. Using month sinusoids was significant almost at 5%. The R squared were 1.15% and 1.05%.

For using month and categorical variables, the R-squared was 4.36 %, it was better than sinusoids but it didn't make a real difference. In our analysis in

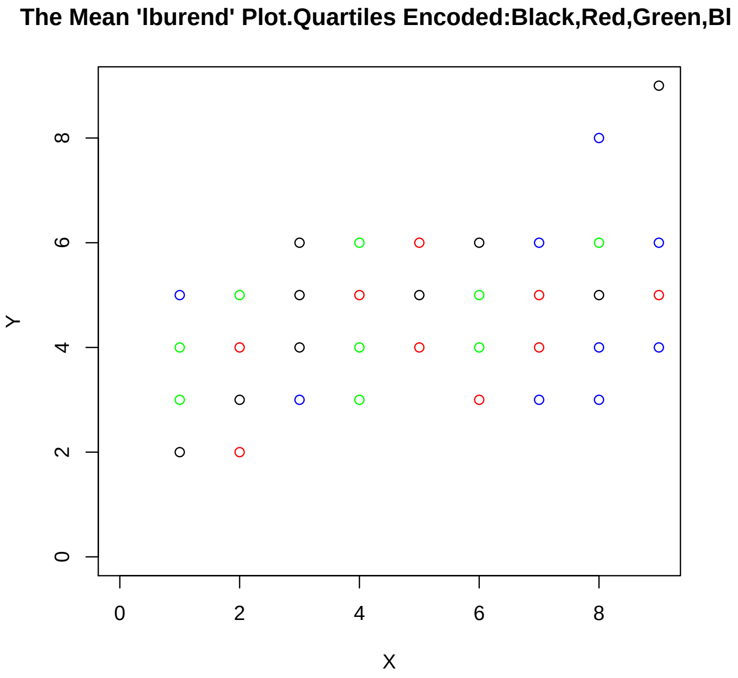

Figure 4.3: The Mean lburend Plot

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesT	4	11.56	2.89	1.48	0.2058
Residuals	512	997.54	1.95		

Table 4.7: The result of sinusoids

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesTmonth	2	10.62	5.31	2.73	0.0659
Residuals	514	998.48	1.94		

Table 4.8: The result of sinusoids just for month

section 4, we are going to use temporal variables for sinusoids for each month and day instead of categorical varables, since the cross validation may fail when using the categorical variables.

Next, we perform exploratory analysis for four FWI components (Fires) and four weather variables. In Table 4.10, it shows the exploratory analysis for four Fire variables (four FWI components). In Table 4.1, it shows the exploratory analysis for four weather variables. They were not statistically significant as shown in the tables and the R-squared were low as 0.8% and 1.04%.

Now we look at other groups of variables considered in Cortez and Morais For the feature selection setup STFWI (X and Y coordinates for spatial, month and day categorical variables for temporal, and four FWI components), only month12 and DMC variables are important. The overall p-value is 0.1042 and they were not statistically significant. For the feature selection setup STM (X and Y coordinates for spatial, month and day categorical variables for temporal, and four weather variables), only month12 and temp variables are important and the overall p-value was 0.08863. They were not statistically significant, either. However, we replaced the spatial varialbes with the splines, fitted the regression and the overall p-value was improved to p=0.0003547 for STFWI and p=0.0004487 for STM. They are now statistically significant when comparing to the variables used in Cortez and Morais. The splines for spatial variables X and Y improved the models hugely. STFWI and STM became statistically significant by using the modified models of STFWI and STM. The modified models using the improved formulation on spatial variables of X and Y for the feature selection setups STFWI and STM are improved than the models in the Cortez and Morais paper.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
month	11	37.37	3.40	1.76	0.0589
day	6	6.65	1.11	0.57	0.7520
Residuals	499	965.09	1.93		

Table 4.9: The result of month and day categorical variables

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesF	4	8.12	2.03	1.04	0.3869
Residuals	512	1000.98	1.96		

Table 4.10: The result of four FWI components

4.2 Multiple Linear Regression

First, We fit a Multiple Linear Regression on forest fires data. Multiple Linear Regression (MR) was applied on each feature selection setup, STFWI, STM, FWI, M, SF, SM, SFM, and S and predicted the 'lburned' (log transformed) area of forest fires. All attributes except the response variable were standardized to a zero mean and one standard deviation.

For the input variables, we used the splines that we obained previously in section 4.1 instead of X and Y coordinates for spatial variables and sinusoids instead of the categorical variables for temporal variables. We have applied a 10-fold cross validation with 30 replications to each configuration to access the predictive performances. The cross validation may fail when using categorical variables, so that's why we used formulation with days and weeks as sinusoids. This seems preferable to alternative of just deleting.

The Multiple Linear regression models were fitted using lm function in R. The MR parameters were optimized using a least squares algorithm. Multiple Linear Regression (MR) was applied on each feature selection setup using 10-fold cross validation. Then the overall performace was computed using the Mean Absolute Deviation (MAD) and Root Mean Squared (RMSE).

We will compare the error criterion for each data mining method and feature selection setups using the log transformed 'burned' variable (ie. log(x+1)). The MAD and RMSE for Multiple Linear Regression models are provided in Table 4.11.

The MR with S produced the lowest error for RMSE which is the best DM method for RMSE. S is the feature selection setup that contains the seven splines for spatial variables X and Y. We explored that all S variables are statistically significant in section 4.1 earlier. These input variables are important for predicting forest fires. For MAD, the MR with STF provided the lowest error among all multiple linear regression methods and this is the second best data mining method for MAD. All MR methods contain S variables perform better than the feature selection setups F and M which do not contain the splines variables. All MR methods except the feature selection setups F and M provided better MAD errors than all other DM methods except all SVM feature selection setups. Using the modifications on spatial variables, S, provided improvements on mulitple linear regression model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
firesM	4	10.50	2.62	1.35	0.2519
Residuals	512	998.61	1.95		

	LM.MAD	LM.RMSE
STF	1.13291	1.39078
STM	1.13759	1.41969
SF	1.13608	1.39322
SM	1.14265	1.42473
SFM	1.14240	1.43302
S	1.13338	1.38741
F	1.16011	1.40264
M	1.17472	1.44003

Table 4.11: MAD and RMSE for Multiple Linear Regression

The summary result of MR with S that we obtained after running a 10-fold cross validation is shown in Table 4.12 and the result of MR with STF is shown in Table 4.13. The both linear regression models are statistically siginificant. The overall p-value for S was 0.0015 and the overall p-value for STF is 0.00065. The p-value of the MR with STF was a little less than the MR with S. We'll later compare both of the MR models for the REC curves with other DM models in secton 4.8. We will compare the MR model with other Data Mining methods.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0899	0.0632	17.24	0.0000
X1	-0.3450	0.0872	-3.95	0.0001
X2	0.1740	0.0975	1.78	0.0750
X3	-0.2204	0.0933	-2.36	0.0186
X4	0.1414	0.1067	1.33	0.1858
X5	-0.3090	0.1029	-3.00	0.0028
X6	0.0851	0.1048	0.81	0.4174
X7	0.1853	0.0961	1.93	0.0545
Y1	0.7278	0.2928	2.49	0.0133
Y2	0.3254	0.1381	2.36	0.0189
Y3	3.7520	1.0316	3.64	0.0003
Y4	-4.2569	1.2269	-3.47	0.0006
Y5	2.7842	0.6987	3.99	0.0001
Y6	0.4134	0.3250	1.27	0.2040

Table 4.12: The results of MR with S

4.3 Neural Networks

Next, we fitted Neural Netsworks on forest fires data. The same preprocessing was used as the previous method (MR) such as using indicator variables and standardization. The neural network is fitted using the nnet function in library nnet in R. We will fit the feed-forward neural networks in which inputs are connected to one or more nodes in the input layer, and these nodes are connected forward to further layers until they reach the output layer. The input nodes are used to represent the input attributes and an ouput node is used to represent the model output. The input nodes are connected forward to each and every node in the hidden layer, and these hidden nodes are conneted to the single node in the output layer. We will consider multilayer perceptrons with one hidden layer of H hidden nodes and logistic activation functions and one output node with a linear function. The jth node of the hidden layer of the feed-forwad network [6] is

$$h_j = f_j(\alpha_{0j} + \sum_{i->j} w_{ij}x_i)$$

where x_i is the value of the ith input node, $f_j(.)$ is an activation function which is logistic function in here $f_j(z) = exp(z)/1 + exp(z)$. α_{0j} is called the bias, the summation i->j means summing over all input nodes feedting to j, and w_{ij} are the weights.

For the output layer, the node is defined as

$$o = f_o(\alpha_{0o} + \sum_{j->o} w_{jo}h_j),$$

where the activation function $f_o(.)$ is either linear or a Heaviside function.

If the output activation function is linear, then the output of a feed-forward neural network (FFNN) can be written as

$$o = \alpha_{0o} + \sum_{j=1}^{k} w_{jo}h_j,$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0914	0.0627	17.41	0.0000
X1	-0.3297	0.0879	-3.75	0.0002
X2	0.1906	0.0981	1.94	0.0528
X3	-0.2116	0.0934	-2.27	0.0239
X4	0.1904	0.1078	1.77	0.0780
X5	-0.3298	0.1030	-3.20	0.0015
X6	0.1340	0.1049	1.28	0.2019
X7	0.2574	0.0977	2.63	0.0087
Y1	0.8220	0.2932	2.80	0.0053
Y2	0.2996	0.1383	2.17	0.0309
Y3	3.8789	1.0273	3.78	0.0002
Y4	-4.4127	1.2220	-3.61	0.0003
Y5	2.8693	0.6960	4.12	0.0000
Y6	0.3214	0.3249	0.99	0.3232
month1	0.2817	0.0852	3.31	0.0010
month2	-0.4779	0.2600	-1.84	0.0668
day1	-0.0945	0.0646	-1.46	0.1444
day2	-0.0262	0.0641	-0.41	0.6832
FFMC	0.1263	0.0771	1.64	0.1020
DMC	0.1844	0.1098	1.68	0.0937
DC	-0.4929	0.2887	-1.71	0.0884
ISI	-0.0478	0.0784	-0.61	0.5419

Table 4.13: The results of MR with STF

Hidden Nodes	STFWI	STM	FWI	M	SF	SM	SFM	S
NN	4	6	4	4				
NN which we obtained	2	2	2	2	2	2	2	2

If the output activation function is linear, then the output of a skip-layer feed-forward neural network can be written as

$$o = \alpha_{0o} + \sum_{i=1}^{l} \alpha_{io} x_i + \sum_{j=1}^{k} w_{jo} h_j,$$

where the first summation is summing over the input nodes, l is the number of input nodes, k is the number of nodes in the hidden layer and h_j is given above. The second equation allows the direct connections from the input layer to the output layer which is referred as a skip-layer feed-forward network.

The NN performance will depend on the value of H. The best hidden nodes for each feature selection setups are as above. A internal 10-fold grid search was used to find the best H. After selecting the H value, the NN model was retrained with all training data.

Using the best hidden node, the Neural Net model is fitted for each feature selection setups with all training data in R. E = 100 epochs is used.

We fitted both FFNN and a skip-layer FFNN for each feature selection setups using 10-fold cross validation. The thirty runs of a 10-fold cross validation (in a total of 300 simulations) were applied to each tested configuration.

For the feature selection, STFWI, we obtained a 21-2-1 network with 47 weights for NN and with 68 weights for a skip-layer NN. For STM, we obtained a 21-2-1 network with 47 weights for NN and with 68 weights for a skip-layer NN. For FWI, a 4-2-1 network is obtained with 13 weights for NN and with 17 weights for a skip-layer NN. For M, a 4-2-1 network is obtained with 13 weights for NN and with 17 weights for a skip-layer NN. We obtained the estimates of their biases and weights using BFGS algorithm using nnet function in R.

After a feed forward neural network is built, it is used to predict the lburned area of forest fires for each feature selection setups. Then the MAD and RMSE are computed in Table 4.14 and Table 4.15.

The NN with S provided the lowest error among all NN methods for MAD. A skip-layer NN with S provided lower error than NN with S for MAD. These input variables, the splines for spatial variabes, are important for predicting forest fires. Also, the NN with M provided the lowest error among all NN methods for RMSE. For S, a 13-2-1 network is obtained with 31 weights for NN and with 44 weights for a skip-layer NN.

	NN.MAD	NN.RMSE
STF	1.19010	1.51819
STM	1.20304	1.52553
SF	1.18761	1.48875
SM	1.18012	1.47415
SFM	1.20802	1.55030
S	1.15458	1.43172
F	1.17589	1.43711
M	1.15935	1.41969

Table 4.14: MAD and RMSE for NN

	NNSkip.MAD	NNSkip.RMSE
STF	1.23189	1.55715
STM	1.23178	1.59109
SF	1.19442	1.51205
SM	1.20097	1.55867
SFM	1.22612	1.58248
S	1.15499	1.44155
F	1.17411	1.43631
M	1.16536	1.51309

Table 4.15: MAD and RMSE for a skip layer NN

4.4 Support Vector Machine

Next, we fit a Support Vector Machine for each feature selection setups for forest fires data. We fit the Support Vector Machine (SVM) using svm function in e1071 library in R. This library uses LIBSVM version 2.88. We used the same input variables as they are used in the previous data mining methods.

In SVM regression, the input x $\in \mathbf{R}^A$ is transformed into a high m-dimensional feature space, by using a nonlinear mapping. Then, the SVM finds the best linear seperating hyperplane in the feature space: [4] [3]

$$\hat{y} = b + \sum_{i=1}^{m} w_i \phi_i(x)$$

where $\phi_i(x)$ represents a nonlinear transformation, according to the kernel function $K(x,x') = \sum_{i=1}^{m} \phi_i(x)\phi_i(x')$. We used the popular Radial Basis Function kernel, which presents less hyperparameters and numerical difficulties than other kernels (e.g. polynomial or sigmoid), $K(x,x') = exp(-\gamma||x-x'||^2), \gamma > 0$. To estimate the best SVM, the ε-insensitive loss function is used.

Given a training set of instance-label pairs (x_i, y_i), i=1, ..., l, the SVM regression require the solution of the following optimization problem. [2] SVM regression performs linear regression in the high-dimension feature space using ε-insensitive loss and, at the same time, tries to reduce model complexity by minimizing $w^T w$. This can be described by introducing (non-negative) slack variables, to measure the deviation of training samples outside ε-insensitive zone.

Thus SVM regression is formulated as minimization of the following.

$$\begin{gathered} \min\ (1/2)w^T w + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ \text{subject to } w^T\phi(x_i) + b - y_i \leq \varepsilon + \xi_i, \\ y_i - w^T\phi(x_i) - b \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, ..., l. \end{gathered}$$

This optimization problem can transformed into the dual problem, the dual problem can be solved numerically using quadratic programming techniques, and its solution is given by $f(x) = \sum_{i=1}^{n_{sv}}(\alpha_i - \alpha_i^*)K(x_i, x) + b$ s.t. $0 \leq \alpha_i^* \leq C, 0 \leq \alpha_i \leq C$, and n_{sv} is the number of support vectors.

The SVM performance is fitted and affected by three parameters, C which is a trade-off between the model complexity and the amount up to which deviations larger than ε are tolerated, ε which is the width of the ε-insensitive zone, and γ which is the parameter of the kernel. C=3 and $\varepsilon = 3\hat{\sigma}\sqrt{ln(N)/N}$ are used as heuristics proposed in Cortez and Morias [3] citeCM, where where $\hat{\sigma}$ is the standard deviation as predicted by 3-nearest neighbour algorithm.

The SVM is fitted using three parameters, C, ε, and γ and using the Sequential Minimal Optimization algorithm.

The thirty runs of a 10-fold (in a total of 300 simulations) were applied to each feature setups in order to find the best $\gamma \in \{2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$ and the selected γ are shown in table below. In the table, the first row presents the γ obtained from the paper [3] and the second row are the γ we have obtained from our analysis. We obtained a little different best gamma parameters for each feature selection setups comparing to the paper. [3]

γ	STFWI	STM	FWI	M	SF	SM	SFM	S
SVM	2^{-5}	2^{-3}	2^{-3}	2^{-3}				
SVM	2^{-5}	2^{-7}	2^{-9}	2^{-5}	2^{-9}	2^{-5}	2^{-5}	2^{-5}

Then, we applied thirty runs of a 10-fold cross validation for SVM for each feature selections using those parameters, predicted the 'lburned', and computed the overall performace by using MAD and RMSE.

The MAD and RMSE for SVM are shown in Table 4.16. The SVM with STM provided the lowest error for MAD and it is the best DM model for MAD. For MAD, all feature selection setups for SVM performed well on predictions. They provided the lowest errors among all other DM methods for MAD. The paper also claimed SVM is the best method and SVM with M provided the best error for MAD. We obtained SVM with STM provided the best error for MAD and SVM with STM provided slightly better MAD than SVM with M.

The SVM with STM provided the lowest error among all SVM methods for RMSE as well, but SVM-STM for RMSE was not very good. The improved spatial variables also improved the SVM model. Later, we will plot and compare the REC curves for the SVM-STM and the SVM-M to see which performs better.

	SVM.MAD	SVM.RMSE
STF	1.12070	1.47097
STM	1.07333	1.43228
SF	1.09962	1.49069
SM	1.09683	1.44680
SFM	1.11203	1.45390
S	1.12201	1.46422
F	1.10041	1.50999
M	1.07506	1.45794

Table 4.16: MAD and RMSE for all SVM

4.5 Random Forest

We fit a Random Forest on each feature selection setups for forest fires data.

The RF is an ensemble of T unpruned Decision Tree, using random feature selection from bootstrap training samples. The RF predictor is built by averaging the outputs of the T trees. In general, RF exhibits a substantial improvement over a single Decision Tree. Here is the algorithm for Random Forest below. [5]

Algorithm: Random Forest for Regression or Classification.

1. For b = 1 to B:
(a) Draw a bootstrap sample Z^* of size N from the training data.
(b) Grow a random-forest tree T_b to the bootstrapped data, by re- cursively repeating the following steps for each terminal node of the tree, until the mini-

mum node size n_{min} is reached.
i. Select m variables at random from the p variables.
ii. Pick the best variable/split-point among the m.
iii. Split the node into two daughter nodes.

2. Output the ensemble of trees ${T_b}_1^B$

To make a prediction at a new point x:
Regression: $\hat{f}_{rf}^B(x) = 1/B \sum_{b=1}^{B} T_b(x)$.
Classification: Let $\hat{\mathcal{C}}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{\mathcal{C}}_{rf}^B(x)$ = majority vote ${\hat{\mathcal{C}}_b(x)}_1^B$.

We used Random Forest using rpart in R. The default parameters were adopted for the Random Forest. Thirty runs of a 10-fold cross validation were applied to eahch tested configuration, predicted the 'lburned', and then the overall performance is computed using MAD and RMSE. The results of the errors are shown in table 4.17.

The RF with S provided the lowest error among all RF method for both MAD and RMSE. For RMSE, RF performed well on predictions. They predicted the second best among all RMSE. The paper metioned that the RF with M provided the second best DM method for MAD. However, the MR with STF provided the second best DM method for MAD since the modifications on spatial variables improved the model. The RF with S provided better MAD than RF with M. Also, for RMSE, RF with S provided better RMSE than RF with F, which is the lowest error among all RF in the paper [3]. Therefore, the S (splines for spatial variables) made improvements on MAD and RMSE for Random Forest.

	RF.MAD	RF.RMSE
STF	1.18223	1.47842
STM	1.18341	1.48241
SF	1.19472	1.48204
SM	1.16324	1.45164
SFM	1.19352	1.52483
S	1.14457	1.38867
F	1.18916	1.45412
M	1.16569	1.44305

Table 4.17: MAD and RMSE for all RF

4.6 LASSO and LAR

We also applied LASSO and LAR for each feature selection setups for forest fires data.

The Lasso is a shrinkage and selection method for linear regression. In Lasso Regression, given training data $(x_1, y_1, ..., (x_N, y_N)$, the lasso estimate is defined by

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

We can re-parametrize the constant β_0 by standardizing the predictors; the solution for $\hat{\beta}_0$ is $\bar{y}$, and thereafter we fit a model without an intercept.

We can also write the lasso problem in the equivalent Lagrangian form

$$\hat{\beta}^{lasso} = argmin_{\beta} \{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

Therefore, the main difference between Lasso and Ridge is that Lasso measures shrinkage by $\sum_{j=1}^{p} |\beta_j|$ while ridge uses $\sum_{j=1}^{p} \beta_j^2$. For Lasso, this has the interesting and desirable effect of setting coefficients to zero. The latter constraint makes the solutions nonlinear in the y_i and there is no closed form expression as in ridge regression. Computing the lasso solution is a quadratic programming problem, although we see that efficient algorithms are available for computing the entire path of solutions as λ is varied. t should be adaptively chosen to minimize an estimate of expected prediction error.

Least angle regression (LAR) is intimately connected with the lasso, and in fact provides an extremely efficient algorithm for computing the entire lasso path.

Here is an algorithm that provides the details for LAR. [5]

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, $\beta_1, \beta_2, ..., \beta_p = 0$.
2. Find the preictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $< \mathbf{x}_j, \mathbf{r} >$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move $\beta_j and \beta_k$ in the direction defined by their joint least squares coefficient of the current residual on $< \mathbf{x}_j, \mathbf{x}_k >$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After min(N-1,p) steps, we arrive at the full least-squares solution.

In step 5, if p > N-1, the LAR algorithm reaches a zero residual solution after N-1 steps (the -1 is because we have centered the data).

For LASSO, there is a simple modification of the LAR algorithm that gives the entire lasso path, which is also piecewise-linear.

LAR: Lasso Modification.
4a. If a non-zero coefficient hits zero, drop its variable from the active set of

variables and recompute the current joint least squares direction.

This is why the LAR algorithm and lasso start to differ when an active coefficient passes through zero.

LASSO and LAR are fitted using lars function in package lars in R. We use 10-fold cross validation with 30 replications for each LASSO and Lar for each configuration of forest fires data. Then we forcasted 'lburned' using LASSO and LAR, then MAD and RMSE for each configuration are computed and compared in table 4.18 and 4.19. The coefficients we obtained from LASSO with STF using 10-fold cross validation is shown in the table below.

The MAD and RMSE for LASSO and LAR are shown in table 4.18 and 4.19. LASSO and LAR provided similar errors but all LASSO except F and M are better than LAR for MAD. The LASSO and LAR with the feature selection setups F and M are the same. For MAD, the feature selection setups which contain S are better than the ones which do not contain S which are F and M. The LASSO and LAR with STF provided the lowest error among all LASSO and LAR for MAD. The LASSO and LAR with F provided the lowest error among all LASSO and LAR for RMSE. The LASSO with STF provided the third best DM methods for MAD. (The LM wit STF was the second best DM methods for MAD.) The splines for the spatial variables also improved the LASSO and LAR models.

In section 4.8, we will compare LASSO with STF with other methods for REC curves.

	LASSO.MAD	LASSO.RMSE
STF	1.14354	1.40825
STM	1.14743	1.43451
SF	1.14640	1.40702
SM	1.15303	1.43947
SFM	1.15376	1.44799
S	1.14435	1.40304
F	1.16148	1.40262
M	1.16555	1.41062

Table 4.18: MAD and RMSE for LASSO

	LAR.MAD	LAR.RMSE
STF	1.14409	1.40668
STM	1.15039	1.43276
SF	1.15010	1.40760
SM	1.15663	1.43576
SFM	1.15681	1.44463
S	1.14791	1.40345
F	1.16148	1.40262
M	1.16555	1.41062

Table 4.19: MAD and RMSE for LAR

	X1	X2	X3	X4	X5	X6	X7	Y1	Y2	Y3	Y4
LASSOcoef	-0.24	0.14	-0.19	0.17	-0.26	0.14	0.31	0.78	0.19	2.61	-3.01
	Y5	Y6	month1	month2	day1	day2	FFMC	DMC	DC	ISI	
LASSOcoef	2.00	0.00	0.31	-0.27	-0.02	-0.04	0.14	0.23	-0.36	-0.03	

Table 4.20: The coefficients of LASSO with STF

4.7 Summary of Results

We have computed MAD and RMSE errors for all methods that are Naive, LASSO, Lar, NN, a skip-layer NN, SVM, and RF in Table 4.21 and 4.22. The naive average predictor was also added in the first column of the table. The naive predictor was computed by averaging the response variable area, applied thirty runs of 10-fold cross validation for each feature selection setups, and then MAD and RMSE are computed. We have compared log transformed data because it is more reliable and accurate to make the comparisons using the transformed data (ie. log(x+1)). We have found that the SVM with the feature selection STM provided the lowest error for MAD. For MAD, the SVM methods with all feature selection setups provided the smallest errors than other methods. The MAD for all SVM provided less errors than all other methods. Thus, SVM is the best method for MAD. In the Cortez and Morais paper, they also claimed that SVM is the best method for MAD, but they claimed that SVM with M (four weather variables) was the best for MAD. However, we have found that the SVM with STM (splines for spatial, sinusoids for temporal and four weather variables) was best for MAD and it provided a little less MAD than SVM with M. So, the modified STM made the improvement on the model. The Linear Multiple Regression (MR) with STF feature selection was the second best data mining method among all other DM methods for MAD. STF contains splines for spatial, sinusoids for temporal and four FWI variables. The LASSO with STF was the third best DM method for MAD. Thus, the modifications we made on variables improved the model. In paper, they claimed that the second best DM model for MAD was Random Forest with M feature selection. But the Random Forest with S was found to be the fourth best DM method for MAD. The MAD of the RF with S was less than the RF with M. So, the splines for spatial variables are important as it was already seen in the previous sections.

For RMSE, we have found that the LM with the feature selection S provided the lowest error and it was the best method for RMSE. S contains the splines for spatial variables. Random Forest with S provided the second best RMSE as well as the second best DM method for RMSE and the LM with STF and LM with SF are the third and forth best RMSE. Among data mining methods, the third best DM method for RSME was Naive (with SM). All the RMSE for Naive method are simliar for all feature selections, but not same because we applied a 10-fold cross validation and averaged the response variable 'lburend'. (Since Naive-SM provided the best for RMSE among all Naive, we'll plot this with other methods in REC curve.) Then, the fourth and fifth DM method was Lar and LASSO with F and S. The results are in Table 4.22. Therefore, S improved models a lot. In paper [3], they claimed that the Naive method was the best for

RMSE and the RF with F was the second best DM method. S improved the model a lot especially for linear regression. The Random Forest with S provided the lower RMSE (which is better) than RF with F. So, S improved RF model as well.

For Neural Networks, it didn't perform well compare to other DM methods for MAD and RMSE.

	Naive	LASSO	LAR	LM	SVM	NN	NNSkip	RF
STF	1.15997	1.14354	1.14409	1.13291	1.12070	1.19010	1.23189	1.18223
STM	1.15939	1.14743	1.15039	1.13759	1.07333	1.20304	1.23178	1.18341
SF	1.15974	1.14640	1.15010	1.13608	1.09962	1.18761	1.19442	1.19472
SM	1.15934	1.15303	1.15663	1.14265	1.09683	1.18012	1.20097	1.16324
SFM	1.15960	1.15376	1.15681	1.14240	1.11203	1.20802	1.22612	1.19352
S	1.15970	1.14435	1.14791	1.13338	1.12201	1.15458	1.15499	1.14457
F	1.15959	1.16148	1.16148	1.16011	1.10041	1.17589	1.17411	1.18916
M	1.15949	1.16555	1.16555	1.17472	1.07506	1.15935	1.16536	1.16569

Table 4.21: MAD for all methods

	Naive	LASSO	LAR	LM	SVM	NN	NNSkip	RF
STF	1.40039	1.40825	1.40668	1.39078	1.47097	1.51819	1.55715	1.47842
STM	1.39975	1.43451	1.43276	1.41969	1.43228	1.52553	1.59109	1.48241
SF	1.40007	1.40702	1.40760	1.39322	1.49069	1.48875	1.51205	1.48204
SM	1.39966	1.43947	1.43576	1.42473	1.44680	1.47415	1.55867	1.45164
SFM	1.39984	1.44799	1.44463	1.43302	1.45390	1.55030	1.58248	1.52483
S	1.40000	1.40304	1.40345	1.38741	1.46422	1.43172	1.44155	1.38867
F	1.39994	1.40262	1.40262	1.40264	1.50999	1.43711	1.43631	1.45412
M	1.39979	1.41062	1.41062	1.44003	1.45794	1.41969	1.51309	1.44305

Table 4.22: RMSE for all methods

For a more detailed analysis to the quality of the predictive errors, the REC curve is provided in the next section.

The REC curve of SVM, LM, RF, Naive, and LASSO will be plotted and compared.

4.8 Regression Error Characteristic (REC) Curve

REC curves were introduced in ML (Bi and Bennett, 2003) [1] to compare the predictive ability of regression models. This method was used by Cortez and Morais (2007) [3] to compute predictions from multiple linear regression with RF and SVM models. On the basis of the REC curves, Cortez and A. Morais (2007) claimed that the SVM model forecast better than multiple linear regression for small fires.

Let e_i, i=1, ..., n be the residuals in a fitted regression model with n observations. For the plot we use either the absolute residual, ϵ_i=$|e_i|$ or the squared residuals ϵ_i=e_i^2. The ϵ_i are referred to briefly as errors and we assume that they have been sorted in ascending order, $\epsilon_1 \leq ... \leq \epsilon_n$. Then we define the accuracy,

$$a(\epsilon) = i/n,$$

where i is the smallest value such that $\epsilon_i \leq \epsilon$. Then $a(\epsilon)$ is a step function and it is equivalent to the empirical cumulative distribution function for ϵ_i, i=1, ..., n.

In most situations, especially with continuous variables, the ϵ_i are all distinct. In this case the distinct values for the accuaracy are simply $a(\epsilon_i)$=i/n, i=1, ..., n. In the case where there are ties, an adjustment is needed. For example, if we write $\underline{\epsilon}$=$\epsilon_1, ..., \epsilon_n$ and $\underline{\epsilon}$=$\{1, 2, 2, 4, 5\}$ then $a(\underline{\epsilon})$=$\{1/5, 3/5, 3/5, 4/5, 1\}$.

For a postive random variable, it may be shown that its expectation is the area under its cumulative distribution funciton. Hence the area over the curve (AOC) indicates the average model error. When ϵ is the squared residual, AOC indicates the residual variance.

When ϵ represents squared error,

$$R^2 \doteq (AOC_{NULL} - AOC_{MODEL})/AOC_{NULL}$$

where R^2 is the coefficient of determination. The REC curve is used to compare the predictive ability of fitted models. To set a benchmark we may consider the null model which is simply a constant. This means there is no dependence of the inputs. In this case the sample mean or median of the inputs may be used and the corresponding model errors computed.

Here is our plot of REC curve for the DM models in figure 4.4. We will plot the REC curve of SVM-STM, SVM-M, LM-STF, LASSO-STF, LM-S, RF-S, Naive-SM. The SVM-STM is the best method for MAD, LM-STF is the second best DM method for MAD, LM-S is the best method for RMSE, and RF-S is the second best RMSE. LASSO-STF is the third best DM method for MAD and Naive (-SM) is the third best DM method for RMSE. The Naive method is a null model in here. They are plotted in different colors. The REC curve of SVM-STM is plotted in black, SVM-M is in grey, LM-S is in red, RF-S is in green, LM-STF is in pink, LASSO-STF is in light blue, and Naive-SM is in blue. SVM-M is the second best method for MAD and we will include this curve in figure 4.4 to compare with the result in the paper [3], since they claimed that the SVM-M is the best DM method for predicting forest fires.

From figure 4.4, the SVM-STM (black) and LM-STF (pink) predicted the 'lburned' of the forest fires well. The SVM with STM (black) and SVM with M (grey) produced the similar REC curves. The SVM-STM was a little above the SVM-M but they are almost the same. SVM is surpassing LM-S if absolute error between 0 and 1.4 allowed. LM-S is surpassing the SVM after absolute error of 1.5 is allowed. LM-STF is surpassing the SVM after absolute error of 1.6 is allowed. Regarding the native predictor, it is the worst method. It predicted the lowest percentage of examples, which is 29% of the examples (log transformed), if an error of 1 is accepted. It was surpassing SVM after absolute error of about 1.3.

The SVM with STM and the SVM with M were the best if absolute error of about 1 is allowed and the LM with S and STF performed best after absolute error of about 2 is allowed. However, LM with STF was better than LM with S for REC curves. The LM with STF with pink color is above the LM with S with

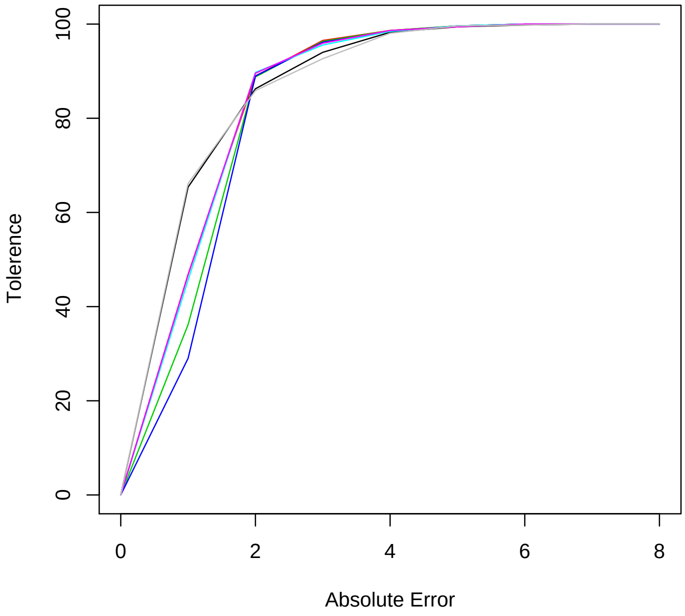

Figure 4.4: REC curves: SVM-STM (black), SVM-M (grey), LM-S (red) , RF-S (green), LM-STF (pink), LASSO-STF (light blue), and Naive (-SM) (blue)

red color. The SVM with STF and the SVM with M performed similarly. Thus, the SVM with STM, the SVM with M and LM with STF are good for predicting forest fires. The paper also mentioned that the SVM predicted well for small fires. In Figure 4.4, the SVM with STM and the SVM with M predicted well for small fires. Thus, the SVM with STM and the SVM with M are the best DM methods if absolute error of about 1 is allowed and the LM with STF is the best method if absolute error of about 2 and more are allowed.

4.9 Other Attempts: Logistic Regression

Other attemps such as fitting Logistic Regression and Multi-response Logistic Regression are presented. We approach this problem as a classfication problem using Logistic Regression and Multi-response Logistic Regression.

4.9.1 Logistic Regression

The logistic regrssion model model arises from the desire to model the posterior probabilities of the K clases ia linear functions in x, while at the same time ensuring that they sum to one and remain in [0,1]. The model has the form [5]

$$log(Pr(G = 1|X = x)/Pr(G = K|X = x)) = \beta_{10} + \beta_1^T x$$

$$log(Pr(G = 2|X = x)/Pr(G = K|X = x)) = \beta_{20} + \beta_2^T x$$

$$.$$

$$.$$

$$log(Pr(G = 2|K - 1)/Pr(G = K|X = x)) = \beta_{(K-1)0} + \beta_{K-1}^T x.$$

The model is specified in terms of K-1 log-odds or logit transformations (reflecting the constraint that the probabilities sum to one).

A simple calculation shows that

$$Pr(G = k|X = x) = exp(\beta_{k0} + \beta_k^T x)/(1 + \sum_{l=1}^{K-1} exp(\beta_{l0} + \beta_l^T x)), k = 1, ..., K - 1,$$

$$Pr(G = K|X = x) = 1/(1 + \sum_{l=1}^{K-1} exp(\beta_{l0} + \beta_l^T x)),$$

and they clearly sum to one.

When K=2, this model is simple, since there is only a single linear function. It is a logistic regression with a binary response.

We fitted Logistic Regression with binary response (two classes). We have two class that are 0 and 1. All y ("burned") values equal to zero are coded as 0 and all y values greater than 0 are coded as 1 The burned variable equal to zero value means that an area lower than $1ha/100 = 100m^2$ was burned. In the dataset, there are 247 samples with a zero value.

We fit a logistic regression with FIRES dataframe that contains splines for spatial variables X and Y. The summary of results are provided in Table 4.24. We obtained AIC: 709.44. An overall test which is likelihood ratio test is performed to test null hypothese. H0: null model (all parameters not significant) vs. H1: fitted model is significant (overall). We obtained p-value=0.001102, so the logistic regression is significant at 0.11% level. The conditional misclassfication rate for 0 is 0.5465587 and for 1 is 0.2481481. The overall misclassfication rate is 0.3907157 and the confusion matrix is shown in Table 4.23. The Logistic Regression with splines predicts class 1 (some fires) slightly better than class 0

	0	1
0	112	135
1	67	203

Table 4.23: Confusion matrix for Logistic Regression using splines for spatial variables.

Then, we fitted a logistic regression with binary response using all input variables. The dataframe x contains all variables that are splines for X and Y, sinusoids for month and day, 4 fire codes, and 4 weather variables. We obtained AIC: 720.8 and an overall test is performed. The overall p-value is 0.005059.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1223	1.0397	0.12	0.9064
X1	-0.5017	0.1322	-3.80	0.0001
X2	0.0254	0.1383	0.18	0.8545
X3	-0.1619	0.1290	-1.26	0.2094
X4	0.0140	0.1566	0.09	0.9286
X5	-0.3473	0.1644	-2.11	0.0347
X6	0.2746	0.1727	1.59	0.1118
X7	0.2862	0.2145	1.33	0.1821
Y1	0.8850	0.5405	1.64	0.1016
Y2	0.7423	0.2040	3.64	0.0003
Y3	10.7271	387.2216	0.03	0.9779
Y4	-12.2130	460.4876	-0.03	0.9788
Y5	7.1442	250.8662	0.03	0.9773
Y6	2.0865	90.7011	0.02	0.9816

Table 4.24: Results from a logistic regression fit to forest fires data using splines data.

The logistic regression is almost statistically significant at 0.5% level. The conditional misclassfication rate for 0 is 0.4939271 and for 1 is 0.2851852. The misclassfication rate is 0.3849130 (a little improved than the previous logistic regression) and the confusion matrix is as follows.

	0	1
0	125	122
1	77	193

Table 4.25: Confusion matrix for Logistic Regression using all x variables.

Next, we fit a logistic regression using step function for the previous model. We obtained AIC: 699.09 and the overall p-value=2.981e-05 which is very small. The logistic regression is statistically significant. The conditional misclassfication rate for 0 is 0.5222672 and for 1 is 0.2703704 The misclassfication rate is 0.3907157 which is same as the first model using splines and the confusion matrix is shown in Table 4.26. The results from the logistic regression is shown in Table 4.27.

	0	1
0	118	129
1	73	197

Table 4.26: Confusion matrix for Logistic Regression using step function.

X1, X5, X6, X7, Y1, Y2, Y3, Y5, month2 and wind variables are selected using step function and all of them are important as shown in Table 4.27.

Logistic Regression predicts class 1 (some fire) slightly better than class 0. The first or third model of the Logistic Regression can be used. However,

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0936	0.0916	1.02	0.3067
X1	-0.4143	0.1129	-3.67	0.0002
X5	-0.2545	0.1241	-2.05	0.0403
X6	0.2953	0.1112	2.66	0.0079
X7	0.2288	0.1226	1.87	0.0621
Y1	0.5180	0.1809	2.86	0.0042
Y2	0.6933	0.1823	3.80	0.0001
Y3	0.4662	0.1420	3.28	0.0010
Y5	0.3581	0.1691	2.12	0.0342
month2	-0.2282	0.0955	-2.39	0.0169
wind	0.1846	0.0946	1.95	0.0511

Table 4.27: Results from a logistic regression fit to all forest fires variables and step function.

predicting the class 0 was not very good (compare to the SVM method. SVM predicted well for small fires). We will fit a Multi-response Logistic Regression in next section and see if they make some improvements than logistic regression.

4.9.2 Multi response Logistic Regression

Now, we fit a Multi-response Logistic Regression.

All y ("burnd") values that are equal to zero are coded as 0. For remaining data (about 250 values) take the quartiles Q1, Q2 and Q3. We observe that Q1=2.14, Q2=6.37, and Q3=15.4225. All y values greater than zero but less and equal to Q1 are coded as 1, all values greater than Q1 but less and equal to Q2 are coded as 2, All values greater than zero Q2 but less and equal to Q3 are coded as 3, and all values greater than Q3 are coded as 4 We use multinom to fit a Multi-response Logistic Regression in package nnet in R. The xy data frame contains splines for spatial variables X and Y and the response y variable. We fitted a multi-response logistic regression using xy data frame and obtained AIC: 1500.739. An overall test which is likelihood-ratio test for multi-logistic is performed. H0: null model (all parameters not significant) vs. H1: fitted model is significant (overall). We obtained p-value=0.01837, so the Multi-response logistic regression is statistically significant at 1.8% level. We computed the misclassfication rate as 0.5125725 and the confusion matrix is shown in Table 4.28. It predicted very well for class 0 which is burned area equal to 0, but not good for other classes. It seems to be not very good.

Next, we fit a Multi response Logistic Regression with all variables that are splines for X and Y, sinusoids for month and day, 4 fire codes, and 4 weather variables. We obtained AIC: 1524.398 which is bigger than previous model and the overall p-value was obtained as 0.001344. The multi-response logistic regression is significant at 0.1% level. We computed the misclassfication rate and it was 0.4893617 which is a little improved than previous model and the confusion matrix is shown in Table 4.29. The predicted values are slightly improved than the previous multi-response logistic regression but it seems to be not very good either.

	0	1	2	3	4
0	246	0	0	0	1
1	66	2	0	0	1
2	66	0	0	0	0
3	64	1	0	0	2
4	63	1	0	0	4

Table 4.28: Confusion matrix for Multi response Logistic Regression using splines.

	0	1	2	3	4
0	231	4	6	4	2
1	51	16	1	0	1
2	54	4	6	0	2
3	56	3	3	2	3
4	54	4	1	1	8

Table 4.29: Confusion matrix for Multi response Logistic Regression using all x variables.

Now, we fit a Multi response Logistic Regression using step function. AIC: 1475.101 and overall p-value=0.0001028. (very small.) The Multi response Logistic Regression is statistically significant. We computed the misclassfication rate as 0.5125725 which is same as the first model multi response logistic regression model and the confusion matrix is shown in Table 4.30. The predicted class is similar to the first model and it doesn't seem to be very good either.

	0	1	2	3	4
0	241	1	1	1	3
1	64	4	0	0	1
2	65	0	1	0	0
3	62	3	0	2	0
4	62	2	0	0	4

Table 4.30: Confusion matrix for Multi response Logistic Regression using step function.

Fitting a Logistic Regression with binary response was better than fitting a Multi-response Logistic Regression. Logistic Regression with binary response provided better misclassification rate than Multi-response Logistic Regression. However, the Logistic Regression predicted well for class 1 but they didn't perform good for class 0. For the Logistic Regression, either the model with spatial variables or the model using step function can be used. However, they predicted well for class 1 but they didn't perform well for predicting forest fires (overall.) The SVM and LM might perform better than Logistic Regression.

Chapter 5

Conclusion

For MAD in table 4.21 in section 4.8, we obtained that the SVM with M is 1.07506 and the SVM with STM is 1.07333. They are similar and the SVM with STM provided a little better MAD by 0.00173. The improved formulation on spatial and temporal variables improved the SVM model. For the REC curves, the SVM with STM and the SVM with M predicted very similarly. The both curves predicted very similar. Also, the paper mentioned that the SVM predicted well for small fires and we obtained the same result from the REC curves in Figure 4.4.

The improved formulation on spatial and temporal variables which is using splines for spatial and sinusiods for temporal improved other models as well. Multiple Linear Regression has improved a lot for using improved saptial and tempral variables than using X and Y coordinates and month and day categorical variables. From MAD and RMSE in table 4.21 and 4.22, we have found that LM with STF is the second best DM method for MAD and and LM with S predicted the best for RMSE criteria. From the REC curve in figure 4.4, it was observed that the SVM with STM and the SVM with M predicted the best if an absolute error of about 1 is allowed and the LM with STF predicted best after an abolute error of about 2 is allowed. Thus, the SVM-STM and the SVM-M are the best methods if an absolute error of about 1 is allowed and the LM-STF is the best method after an abolute error of about 2 is allowed. The result was quite similar to the paper, since the SVM-STM and SVM-M performed simlarly. However, the LM model has improved. For LM, the splines of spatial, the sinusoids for temporal and four FWI components variables are important for predicting forest fires. The seven splines on spatial variables improved the SVM as well. For SVM, the splines of spatial, the sinusoids for temporal and four weather variables are impacting importantly for predicting forest fires now. If a small error tolerence is allowed, then the SVM with STM and the SVM with M predict the best and the LM predicts the best if a little bigger error tolerence is allowed than the SVM.

We fitted a logistic regression and multi-response logistic regression to see if they can make improvements for predicting some fires (greater than 0), since the SVM predicts well for small fires. The logistic regression predicted better for class 1 (some fires) than class 0 (area lower than $1ha/100 = 100m^2$ was burned). However, it didn't predict well for class 0 so it wasn't performing very good.

The modifications made on the spatial and temporal variables improved the models such as MR, SVM, LASSO and LAR, and RF.

The LM with STF outperformed SVM with STM after the absolute error of about 1.6 is allowed and the SVM outperformed LM if absolute error of about 1.5 is allowed. Thus, the SVM-STM and the SVM-M are the best methods if an absolute error of about 1 is allowed and the LM-STF is the best method after an abolute error of about 2 is allowed.

In paper [3], they said that the proposed model which is SVM with M is still useful to improve firefighting resource management. For instance, when small fires are predicted then air tankers could be spared and small ground crews could be sent. Such management would be particularly advantageous in dramatic fire seasons, when simultaneous fires occur at distinct locations. Thus, the SVM with STM may be useful, too. The SVM with STM, the SVM with M, or LM with STF can be suggested for predicting forest fires.

Bibliography

[1] J. Bi and K. P. Bennett. *Regression error characteristic curves.* In Proceedings of the 20th International Conference on Machine Learning, 2003.

[2] Chih-Chung Chang and Chih-Jen Lin. *a Library for Support Vector Machines.* http://www.csie.ntu.edu.tw/∼cjlin/papers/libsvm.pdf, 2009.

[3] Paulo Cortez and Anibal Morais. *A Data Mining Approach to Predict Forest Fires using Meteorological Data.* Cambridge University Press, Department of Information Systems R and D Algoritimi Centre, University of Minho, Portugal, 2007.

[4] Vipin Kumar Pang-Ning Tan, Michael Steinbach. *Introduction to Data Mining.* Addison-Wesley, 2006.

[5] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer, 2009.

[6] Ruey S. Tsay. *Analysis of Financial Time Series.* John Wiley and Sons, Inc., Hoboken, New Jersey, 2005.