A Machine Learning Approach to Better Understand Wildfire-Climate Relationships in the Arctic Tundra (2001-2015)

Arif Masrur
IST 597: Final Project

April 28, 2018

1 Introduction

In recent years different parts of Arctic tundra ecosystems have experienced an increased frequency of wildfire events. Although contemporary studies illuminated on climate-wildfire linkage for regional and historic wildfire events (i.e. in Boreal forests), a detailed understanding is lacking regarding the nature of relationship between climatic parameters and wildfire characteristics (i.e. occurence and intensity) in the Arctic tundra ecosystems. It should be noted that, wildfires exhibit characteristics of complex system dynamics . Studies have found [2] complex, non-linear, intertwined relationships among bio-climatic factors, human influences, and wildfire activity around the planet. In other words, wildfires are largely driven by bottom-up (fuel, weather, and topography) and top-down (climate) controls across a range of spatial and temporal scales. However, as opposed to deriving a detailed understanding of the complex interplay among bottom-up and top-down controls, here I focused on dissecting climatic influences on tundra wildfire activity. Incorporating big bio-climatic datasets for a large-scale analysis is constrained by data availability, as well as by the scope of a class project. Nevertheless, I remained optimistic that a statistical machine learning-based analyses will provide some valuable insights into the climate-wildfire relationships for poorly-understood Pan-Arctic tundra wildfire dynamics.

Objective: the key question answered in this project is: what is the nature of relationship between wildfires (occurrence and intensity) and climate parameters in the Arctic tundra at two temporal scales: monthly and seasonal. Satellite-derived daily wildfire observations since 2000 to date suggest that Arctic tundra wildfire season generally spans over May-October months - an usual summer period in the circumpolar Arctic.

Data: in order to elucidate explicit climate associations of recent Tundra wildfires, I have used *fire occurrence instances (yes/no fire)* and *fire intensity* (measured by radiated heat output) variables in NASA MODIS dataset (2001-2015) as outcome variable, and six climatic parameters in MERRA (Modern-Era Retrospective analysis for Research and Applications) as predictor variables (Table 1). NASA TerraâĂŹs MODIS (Moderate Resolution Imaging Spectroradiometer) daily active fire products (available from 2000 to date) provide a new generation of moderate resolution (1 km) remotely-sensed global fire datasets. On

the other hand, the MERRA-Land dataset contains hourly averages of land surface fields (i.e., surface temperature) from January 1st, 1980 at a horizontal resolution of 2/3° longitude by 1/2° latitude.

MERRA Variable Names	Description	Units
	Explanatory variables (MERRA-Land)	
TSURF	Mean soil **surface temperature** (including snow)	K
PRECTOT	Total **surface precipitation**	kg m^{-2} s^{-1}
SFMC	Top **soil layer moisture** content	kg m^{-2} s^{-1}
EVLAND	**Evaporation** from land	kg m^{-2} s^{-1}
(For detail information on MERRA-Land climate variables, see Gelaro et al, 2017; Reichle, 2012)		
Temperature and Precipitation **Anomalies**	2001-2015 average monthly **TSURF anomaly**	K
	2001-2015 average monthly **PRECTOT anomaly**	kg m^{-2} s^{-1}
	Response variables (MODIS Active Fires)	
Fire instances (yes/no fire)	*Yes* = at least one fire occurred per spatial cell, *No* = no fire occurred per spatial cell (over the study period)	
Avg. FRP	Average **wildfire intensity** per spatial cell	Mega-Watt

Table 1. MODIS active Wildfire and MERRA climate varibales used in this study.

Methods: To investigate wildfire-climate relationships, response variables (wildfire count and wildfire intensity) were aggregated into 2/3° longitude by 1/2° latitude grid cells (conforming to MERRA's spatial scale). Then response and predictor variables (i.e. MERRA climate) were further aggregated at two temporal scales - monthly and seasonal (May-October). All the variables used in two temporal-scale analyses are documented in Table 2.

Seasonal-scale analysis was aimed at identifying if monthly climatic conditions of individual winter (November-February), spring (March-April), and summer (May-October) months had been influential in determining the likelihood of summer season tundra wildfire occurrences between 2001 and 2015. Monthly-scale analysis, in contrast, focused on identifying how average climatic conditions of individual wildfire months and preceding three months (based on fire ecological perspective) had been influential in determining average wildfire intensity aggregated per spatial cells (2/3° longitude by 1/2° latitude). As mentioned previously, wildfire occurrence and intensity are not determined by climatic conditions alone, hence we don't expect that climate variables will be able to explain much variability in the outcomes (i.e. occurrence and intensity). However, with the rapid climate warming in the Arctic and associated vegetation change (i.e. tundra greening) it is reasonable to believe that climatic conditions may have played significant role in wildfire activity in the Pan-Arctic tundra ecosystems over the past 15 years.

For these two temporal-scale analyses following five different machine learning and data mining techniques were used:

1. Naive Bayes,
2. Logistic Regression,

3. Support Vector Machines (SVM),

4. Random Forest, and

5. Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost)

2 Wildfire-Climate Relationship

2.1 Exploratory Data Analysis (EDA)

Because of the time-series and spatial characteristics of wildfire-climate dataset, its essential to explore data distribution of both response and predictor variables. To decide on if response variable for regression (i.e. wildfire intensity) needs any transformation, I regressed the variable `Avg_FRP` on all predictors and look at diagnostics (Appendix 1: Figure 1). Since the residuals from the full model is highly non-normal (i.e. right-skewed), I considered a log-transformation of the response variable - a common solution to positively skewed data. Calling the new response `log.y`, the diagnostics plot for the full model look far better (Appendix 1: Figure 2). No pattern of residuals is detected in the "Residual vs Fitted" plot, and all of them are scattered around 0 with seemingly constant variance. There is a few outliers, but judging from the Residuals vs Leverage, none of them are influential/problematic. Quantile-Quantile plot indicates normality for the majority of residuals with slight departure out in the tail.

1. Linearity and Outliers Diagnosis
   I took a look at the box plot of response against categorical variables Month and Year (Appendix 1: Figure 3). The left plot indicates similarity across all the months from May to October, with similar within group variance, and very small between group variance. The right plot is more informative about differences in fire intensity across the years. For example, 50% of fire in the year 2013 have intensity higher than 75% of fire in most years. Some years (for ex. 2006) had fire intensities that varied much more than others (for ex. 2002, 2010).
   Before discussing linearity, I again looked at the outliers with unusually high/low fire intensity (Appendix 1: Figure 4).I ventured to explain these points as due to lack of variables. Forest fire is a function of Climate, Fuel and Ignition. In the case of fire intensity, Fuel data is even more important. Lacking in variables that capture Fuel and Ignition could be the reason for the weak relationship between response and predictors.
   For most plots, the loess curve and its confidence interval support linear relationship between individual predictors and the response, with the exception of `Lat, Pre3months_Precip`, `Pre3months_Precip_Anomaly`.

2. Autocorrelated Errors Diagnosis
   Since the dataset has potential time trend, I conducted some time-series analysis. First I looked at the plot of `autocorrelated function` and `partial autocorrelated function`. The former is an indicator of autocorrelation of first order, while the latter reveals general stochastic characterization of residuals (Appendix: Figure 5). Contrary to our expectation, the plots in figure 5 suggest that

autocorrelation of residuals are not present. I double checked this with the Durbin-Watson test. The null hypothesis $\rho = 0$ is accepted if the test statistic is close to 2, and is rejected if it is near 0 or 4. Since d = 1.89 (see R output 1) is very close to 2, the test confirmed what the plots have suggested. I could only think of an explanation that perhaps this is due to the non-continuity of the time scale over which data was collected. Fire intensity is only recorded when there is a fire, otherwise no data record is made.

3. Multicollinearity Diagnosis
   In checking with Variance Inflation Factor, after adjusting for the degrees of freedom, none of the predictors exhibited high vif (see R output 2). Despite this, the predictors whose d.f.'s are 1, `Present_Precip, Present_Temp, Present_Precip_Anomaly, Pre3months_Precip, Pre3months_Temp` have vif before adjustment that is greater than 10, especially so in the case of `Pre3months_Temp`. This suggested that multicollinearity is present and perhaps not all predictors will be statistically significant.

2.2 Regression with Transformed Response

`factor.Year.2003`, and `Lat`.

3 Classification

This part focused on analyzing the influence of seasonally aggregated climate parameters for tundra wildfire *occurrence*. Since the response variable is categorical ('yes' or 'no' fire), its a classification problem. Therefore, I started with widely-used classification measure - Naive Bayes and then logistic regression - to model the tundra wildfire occurrence and climate relationship.

$$\Pr(\text{fire} = \text{Yes} \mid 72 \text{ climate parameters})$$

It should be noted that, here I were primarily interested in *prediction*. For both interpretation and prediction, I used tree-based classification measures (e.g. Random Forest) and SVM which are discussed in forthcoming sections.

3.1 Naive Bayes

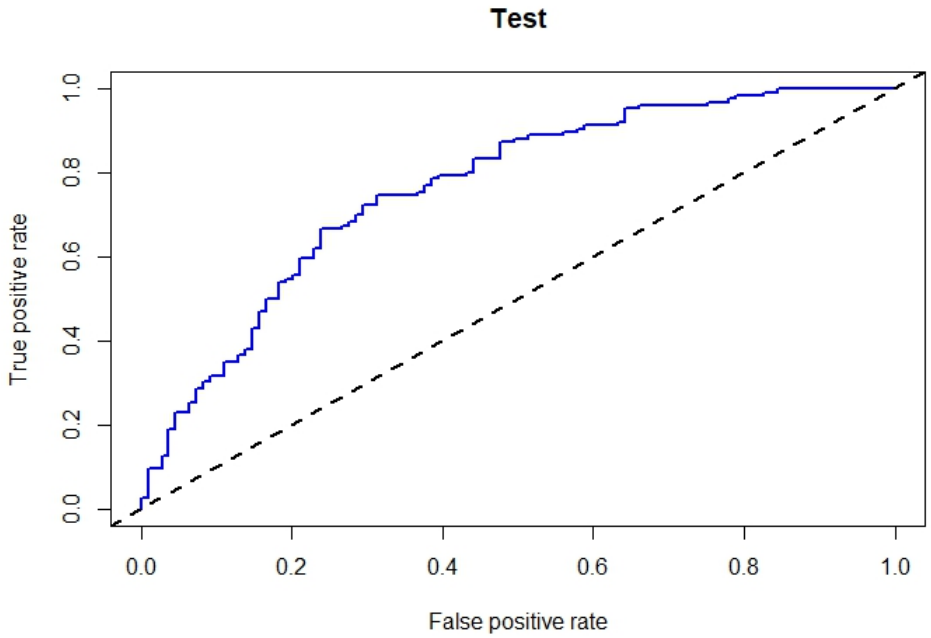

Figure 1. Performance of Naive Bayes Classification.

3.2 Logistic Regression

For this binary logistic regression analysis, the dataset consisted of one response (values are 1 or 0; 'yes' as 1 and 'no' fire as 0) and 72 climatic predictor variables (6 monthly climate variables for 12 months âĂŞ see Table 1 and 2). There are 940 observations (both fire and non-fire cells) among which 704 cells contained wildfire events for each season between 2001 and 2015. The remained 470 non-fire cell observations were randomly chosen around neighboring wildfire cells.

To begin with, I fitted a logistic regression model using glm() function (a class of generalized linear model that includes logistic regression) with binomial link. The output showed only 24 (out of 72) predictors were statistically significant. To address this issue, I performed *over-dispersion* and *collinearity* tests. Pearson residuals were used to check for over-dispersion which allowed analyzing whether the actual variance of response was greater than that suggested by the model.

Pearson residuals appeared to have mean close to zero and variance close to one - an evidence suggestive of no over-dispersion. Additionally, no influential outlier(s) was detected (see R output). In Looking for collinearity, 92 pairs of predictors had correlations greater or equal to 0.7 and 30 pairs had correlation greater or equal to 0.9. To find most relevant variables that are associated with the response ('yes' fire vs. 'no' fire), I used Lasso Regression - a popularly used regularization method that sets coefficient estimates of irrelevant predictors to zero and thereby enhances *model interpretability* and *prediction accuracy* . Additionally, for the purpose of enhancing model's prediction power I used 10-fold cross validation. I randomly partitioned into training (75% of observations) and test (25% of observations) sets.

Values of regularization parameter vs. deviance and coefficients are plotted in (Appendix: Figure 6). The *best lambda* corresponded to CVMSPE values of 1.110754. (see R output 9 for a list of variables and best lambda). I found that the coefficients of 32 predictors were set to zero. This shouldn't come as a surprise given the high collinearity among many predictor variables. The prediction power of the fitted model on test dataset (see R output 10) showed that 71% of times the model correctly predicted the fire occurrences.

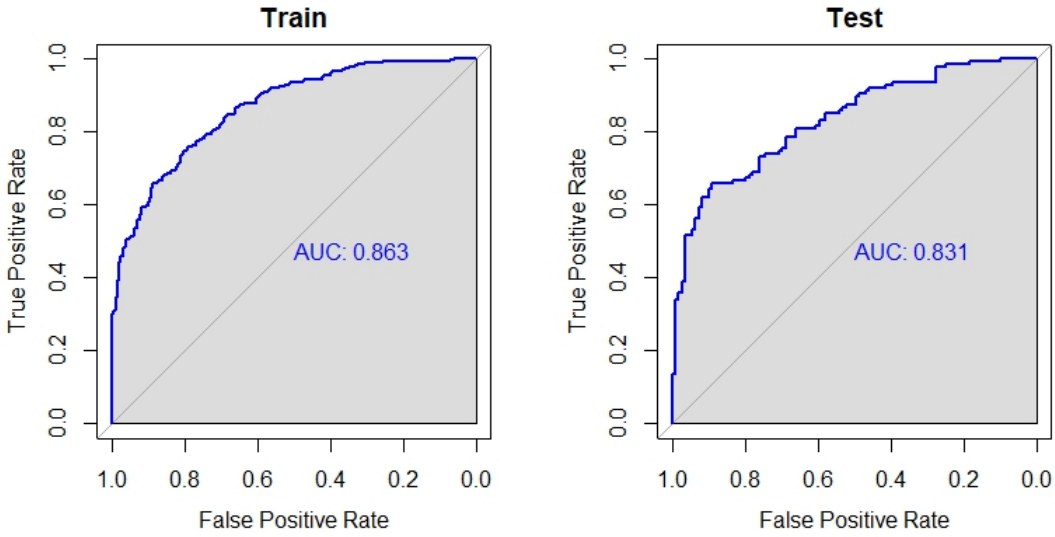

Figure 2. Performance of Logistic Regression Classification.

As mentioned previously that I aimed at finding classification model for prediction of fire occurrences. While logistic model has reasonably well performed in terms of prediction than the Naive Bayes (See AUC values in Figure 1 and 2), this may not hold for unknown future observations. Moreover, in order to be able to make interpretations about coefficients and infer a causal relation between response and predictors further analysis is necessary. Using other methods of regularization, aggregating observations to get 3-month average instead of monthly observations, or using $\acute{\lambda} = \lambda + SD(\lambda)$ as best lambda in lasso regression analysis and Principal Component Analysis are possible approaches towards a model with sensible interpretation.

3.3 Decision Trees: Random Forests

To supplement Logistic Regression, next I focused employed decision tree-based classification method. I chose widely-used (in wildfire research) Random Forest classification - a non-linear, tree-based ensemble learning method, to examine the nature of tundra wildfire-climate relationship, since this method usually yields relatively better classification results among all tree-based methods. As opposed to growing single decision tree (as in CART), random forest grows multiple trees, with having each split to consider only a subset of all predictors.Then it takes average of all trees to make final tree. In this way, random forest can reduce amount of potential correlation between trees and thereby helps reduce the variance of the final tree. In addition, while tree-based method don't have same level of predictive accuracy as classical methods (e.g.

logistic regression), Random Forest can actually yield improved prediction accuracy through a "consensus" prediction based on many aggregated decision trees [4].

Before fitting Random Forest model, I randomly partitioned the wildfire-climate dataset into training (75%) and test (25%) datasets. Therefore, out of total 940 'yes' fire and 'no' fire observations (spatial cells), 705 were randomly designated as training and 235 as test observations. Then, I used tuneRF function to find the optimal numbers of variables to try (known as *mtry*) splitting on at each node. I found mtry = 13 produces least out of the box (OBB) error (14.18%), that means, 13 out of 84 predictors should be considered for each split. Then I fitted random forest model on training cells using mtry = 13 and 500 trees (known as *ntree*). Its worth mentioning that, I also tried other ntree values, e.g. ntree = 100, 1000, 1500, and 2000, which didn't cause any significant differences in model performance. From plotted model (Appendix 1: Figure 7) I can see that, albeit having little fluctuations after about 250 trees there wasn't much changes in terms of prediction error. Black solid line is for overall OOB error, green line represents classification error rate for 'yes' fire, whereas red line represents classification error rate for 'no' fire cases.

I found that the training model fitted the training data fairly well. Out of total 366 'no fire' observations in the training set, 56 of them were misclassified as 'yes' fire observations. On the other hand, out of total 339 'yes' fire observations, 43 of them were misclassified as 'no' fire observations. Therefore, the prediction accuracy was 85.96%. Next, I used this fitted model on training dataset to predict 'yes' fire and 'no' fire instances in the test dataset. The Receiver operating characteristic (ROC) curve for the prediction in test dataset is included in Figure 3.

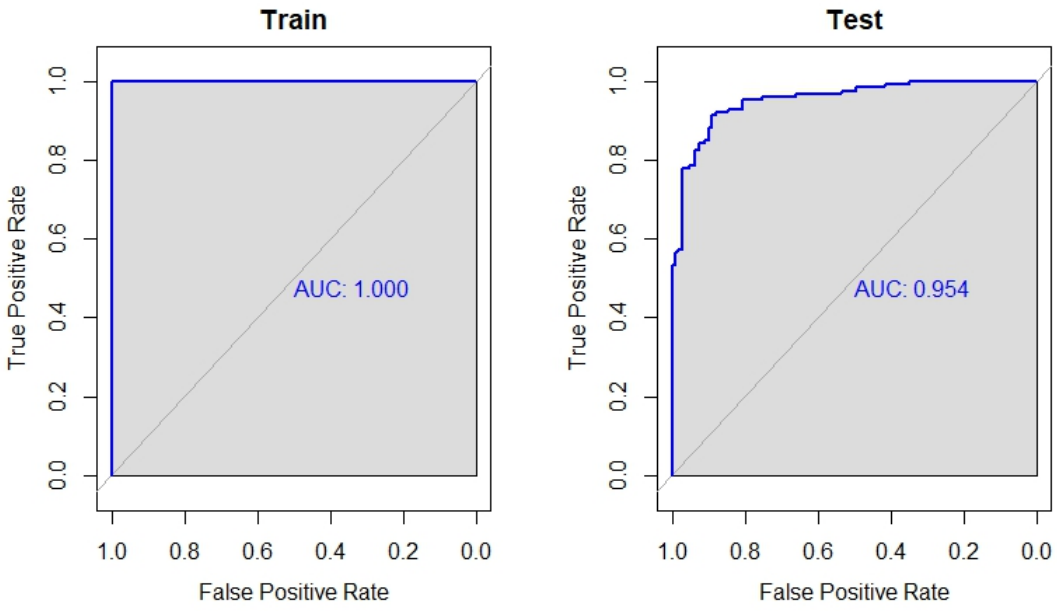

Figure 3. Performance of Random Forest Classification.

In the Variable Importance Plot (Figure 4), predictor variables that resulted in nodes with higher purity have a higher decrease in Gini coefficient. The June surface temperature anomalies variable were by far the most important variable in determining a 'yes' fire vs. 'no' fire instance. Other relatively important climate

variables were surface temperature anomaly of summer (July, October) and winter (February). Also, July surface moisture content may be another important variable. However, it seems overall surface temperature values have been very important drivers for tundra wildfire occurrence in 2001-2015 period.

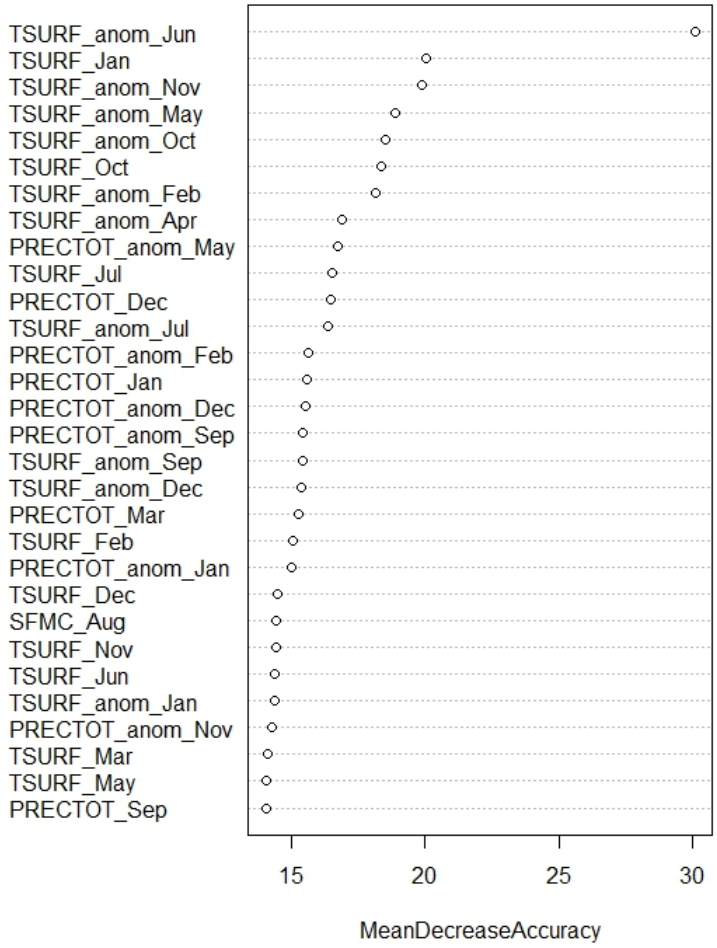

Figure 2. Performance of Random Forest Classification.

3.4 Adaptive Boosting (AdaBoost)

3.5 Extreme Gradient Boosting (XGBoost)

3.6 Support Vector Machine (SVM)

4 Discussion

References

[1] Ades, A. E., et al. âĂIJExpected Value of Sample Information Calculations in Medical Decision Modeling.âĂİ Medical Decision Making, vol. 24, no. 2, 2004, pp. 207âĂŞ227.

[2] Bowman, D.M., Balch, J.K., Artaxo, P., Bond, W.J., Carlson, J.M., Cochrane, M.A., DâĂŹAntonio, C.M., DeFries, R.S., Doyle, J.C., Harrison, S.P. and Johnston, F.H., 2009. Fire in the Earth system. science, 324(5926), pp.481-484.

[3] Strong, Mark, et al. âĂIJEstimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample.âĂİ Medical Decision Making, vol. 35, no. 5, 2015, pp. 570âĂŞ583.

[4] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112). New York: springer.