Predicting Forest Fires in Portugal and Northern Algeria

Abstract

Forest fires are becoming a major, global environmental concern in recent years. Therefore, creating an effective forest fire monitoring system is critical to prevent environmental and economic damage. However, most models created have either included a limited number of environmental factors or not considered interaction effects between environmental factors. In the current study, we performed an analysis on all 11 potential explanatory variables from a merged dataset of Algerian and Portugal forest fire datasets. We used the drop-in-deviance test to conduct a variable selection process on a training dataset to determine which combination of environmental factors best predict the occurrence of a forest fire. Overall accuracy, ROC curves, residual analysis, transformations, and interaction terms were also considered as we developed our model using five of the explanatory variables in our dataset. Overall accuracy in our training dataset is 70.57% while accuracy for our testing data is 65.26%.

Introduction

As climate change and human activities increase, forest fires have increased in intensity and extent throughout the world. Both Algeria and Portugal suffer the consequences of seasonal forest fires, which directly cause destruction of forest ecosystems, loss of lives, and monetary expenses for government agencies (Kutter et al, n.d.; Mateus & Fernandes, 2014).

Research has suggested that forest fire mitigation strategies need to integrate available biological, ecological, physical, and technological fire-related information (Mateus & Fernandes, 2014). Abid and Izeboudjen (2020) produced a fire prediction model with an 82.89% accuracy and 0.92 recall value using a decision tree using Algerian forest fire and found that temperature, relative humidity, and wind speed are important predictors for fire occurrences. Cortez and Morais (2007) used a data mining approach to propose a Support Vector Machines (SVM) model using only four direct weather inputs (temperature, rain, relative humidity, and wind speed) to predict small fires in Montesinho natural park in Portugal.

To investigate what combination of factors best predicts forest fires, we merged two forest fire datasets collected from Algeria and Portugal. The Algeria dataset included 244 instances of forest fire data from two regions of Algeria, namely the Béjaïa region and the Sidi Bel Abbès region from June 2012 to September 2012 (Cortez & Morais, 2007). We removed one observation due to an error within the Algeria dataset. The Portugal dataset included 517 entries of forest fire data from the Montesinho Natural Park located in the Tr´as-os-Montes northeast region of Portugal from January 2000 to December 2003 (Abid & Izeboudjen, 2020). The common variables between these two datasets are region, month, temperature, humidity, rain, and multiple weather indices from the Forest Fire Weather Index system. The Fine Fuel Moisture Code (F Moisture) is one of those indices and represents the moisture content of litter and fine fuels in a forest, and requires temperature, relative air humidity, wind speed, and precipitation as input data (Van Wagner 1987). The Algeria dataset did not measure the area or the intensity of the fire. Therefore, we used only the binary variable of whether or not a fire occurred on that day (BinaryFire) as the response variable.

Methods

Our dataset was split into training and testing subsets before we began our analysis. 70% of our data were in the training dataset, and the remaining 30% were in the testing dataset. In order to account for overfitting, the 70-30 split was used to train on less data from the already small dataset. We conducted a logistic regression analysis to best predict if there is a fire or not on the 11 variables including all two-way interaction terms. We proceeded our analysis with regression subset selection using an exhaustive search using the all-interaction model. In order to compare the models with varying terms, we selected models with low Akaike's Information Criteria (AIC) and Bayesian Information Criterion (BIC) values as the best model. These criteria are used for regression model selection and account for the tradeoff between model complexity and fit. Models with low BIC and AIC values are generally preferred. We used accuracy and deviance as other metrics to compare models. Drop-in-deviance tests were performed to compare models with close AIC and BIC values.

Results

Using our training dataset, we created no-interaction term and all-interaction term models. A drop-in-deviance test comparing the all-interaction model with the no-interaction model suggested that at least one interaction was important and should be included in the model.

Adding all interactions resulted in improvements in various metrics in our model (Appendix 1). Based on the best regression subset selection analysis, the models with 5, 6, and 7 terms had comparably low AIC and BIC values. Since our goal is to minimize forest fires, the model should minimize the number of false negatives. That is, our model should limit the number of times we predict no fire when there actually is a fire. To do this, we used a threshold of 0.4 in our confusion matrix based on the ROC curve and the goal to maximize the true positive fraction, which is a cautious approach that allows for minimal false negatives (Appendix 2). Based on this threshold, we employed a confusion matrix for each model in order to compare accuracy and number of false negatives (Table 1, Appendix 5-8).

Model	AIC	Deviance	Accuracy	False Negatives	False Negative Percent
5 terms (train)	535.52	515.52	70.57%	7	0.01%
6 terms (train)	534.14	508.14	70.38%	11	0.02%
7 terms (train)	534.06	506.06	70.57%	15	0.03%
5 terms (test)	223.82	203.82	65.26%	3	0.01%

Table 1. Evaluation metrics for the 5-term, 6-term, 7-term logistic regression models. Terms for each model can be found in Appendix 10.

The 5, 6, and 7 term models have similar interaction terms included and comparable AIC values, deviance, and accuracy (Table 1, Appendix 10, Appendix 11). Therefore, we decided to conduct the drop-in-deviance test between these three models and the all-interaction model (Appendix 3). The drop-in-deviance test determined that the 5 term model adequately explains the variance in the model (Appendix 3). When the 5 term model was compared to the all-interaction term model there were no noticeable improvements in our metrics, and a drop-in-deviance test showed no significant difference (Appendix 3, Appendix 4). Since this reduced-interaction model balanced predictive power with simplicity, while not compromising accuracy and false negatives, we chose to use the 5 term model, with the addition of all variables shown to have significant interaction terms, as our best overall model. When we evaluated this model with our testing data, we found a small decrease in the overall accuracy and deviance, while the percent of false negatives remained the same (Table 1, Appendix 9). This small drop in accuracy provides evidence that our model based on estimation could be used to predict future datasets while maintaining accuracy.

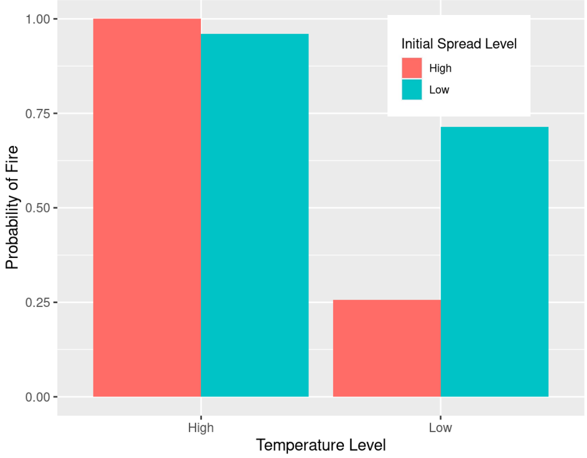


Figure 1. Probability of fire by temperature level and initial spread level (Temperature Level: <=15ºC is Low, >=30ºC is High; Initial Spread Level: >=15 is High, <15 is Low).

Our final 5 term model indicated that temperature is a significant predictor of forest fires, in addition to interactions between temperature, humidity, F moisture, initial spread, and Portugal (Appendix 12). Initial Spread is an index variable calculated through wind speed and F. Moisture that represents the rate of fire spread without the influence of fuel (Wagner, 1987). The interaction effect of temperature and initial spread on fire occurrences indicate that when temperature level is high, high and low rate of fire spread influence the probability of fire equally, but when temperature levels are low, a low rate of fire spread has a higher probability of fire (Figure 1). Our model also shows that Initial spread interacts with Portugal, revealing that region can affect which terms should be included in the model.

Discussion

Our reduced final 5 term model offers an improved model compared to the no-interaction term model and the all-interactions term model. Thus, we created a model to predict forest fire occurrences in Portugal and Northern Algeria using a relatively small number of variables. Although our goal was not to find the significance of individual variables on forest fire occurrences, the overall trend that temperature and humidity are important predictor variables for fire occurrence is corroborated in the current study (Cortez & Morais, 2007; Abid & Izeboudjen, 2020). Compared with the predictive model using the Algeria dataset which had an 82.89% accuracy and 0.92 recall value (Abid & Izeboudjen, 2020), our model does have a lower accuracy of 70.57%. This can be explained by the addition of the Portugal dataset to the current study, which may have made it harder to generate a highly accurate model for two regions with different climate environments. However, our model does have a higher recall value $(\frac{number\ of\ true\ positives}{(number\ of\ true\ positives)*(number\ of\ false\ negatives)})$ at 0.97 compared to 0.92. The number of correctly predicted fires is most important to the current study's aim, which can be done by limiting the number of false negatives and maximizing the number of true positives. The current study's model is further supported by ecological knowledge that daily weather and long-term climate influence fire ignition potential, behavior, and severity – including moisture content (Benson et al., 2008). A previous study that found low humidity and high temperature are both primary factors causing forest fires, which is consistent with the current model's interaction term between temperature and humidity (Varela et al., 2020).

These results must be taken in context. Since our model is limited to data from three specific regions, this model can only be used as predictors for fires in these regions, and therefore has limited generalizability. The interaction term between Portugal and initial spread as a predictor in our model suggests the combination effect of region and initial spread on the fire occurrences. Since region does impact the model, we suggest that region will be an important fire predictor. This could imply that models should be created on a case-by-case basis in order to have the most accurate predictions by region. Additionally, the best model does use both weather measurements along with indices, which means to use the model, one must have all those measurements and calculations. Our addition of weather indexes and interaction terms, which have not been discussed in current literature, may provide insight into the existing fire predicting models, helping inform and protect nature reserves, animals, plants, and people from deadly and destructive fires. The model's limitations demand addressing before it is used in real-world settings to assist with forest fire prevention.

Before any generalizations toward forest fire occurrences in other regions can be made, further research outside of Algeria and Portugal are required. Furthermore, an important aspect of the forest fire – the intensity of the fire – was not addressed in our study, even though the Portugal dataset did measure the fire area. The indicators of the severity of the fire, like the specific fire behavior and the effect of the fire on the landscape could potentially be monitored for future studies as outcome variables. By including fire as a quantitative variable, more meaningful connections between the severity of the fire and environmental factors can possibly be discovered.

References

Abid, F. & Izeboudjen, N. (2020). *Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm.* doi.org/10.1007/978-3-030-36674-2_37.

Benson, R. P., Roads, O. J., & Wiese, D. R. (2008). Climatic and Weather Factors Affecting Fire Occurrence and Behavior. In A. Bytnerowicz, M. J. Arbaugh, A. R. Riebau, & C. Andersen (Eds.), *Developments in Environmental Science* (pp. 37-59). Elsevier. https://doi.org/10.1016/S1474-8177(08)00002-8

Cortez, P. & Morais A. (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Retrieved from: http://archive.ics.uci.edu/ml/datasets/Forest+Fires

Kutter, A., Bih, K. B., & Jauffret, S. (n.d.). *Sustainable forest management will help Algeria tackle the risk of wildfires*. World Bank Blogs. Retrieved December 5, 2022, from https://blogs.worldbank.org/arabvoices/sustainable-forest-management-will-help-algeria-tackle-risk-wildfires

Mateus, P., Fernandes, P.M. (2014). Forest Fires in Portugal: Dynamics, Causes and Policies. In F. Reboredo (Eds.), *Forest Context and Policies in Portugal* (pp. 97-115). Springer, Cham. https://doi.org/10.1007/978-3-319-08455-8_4

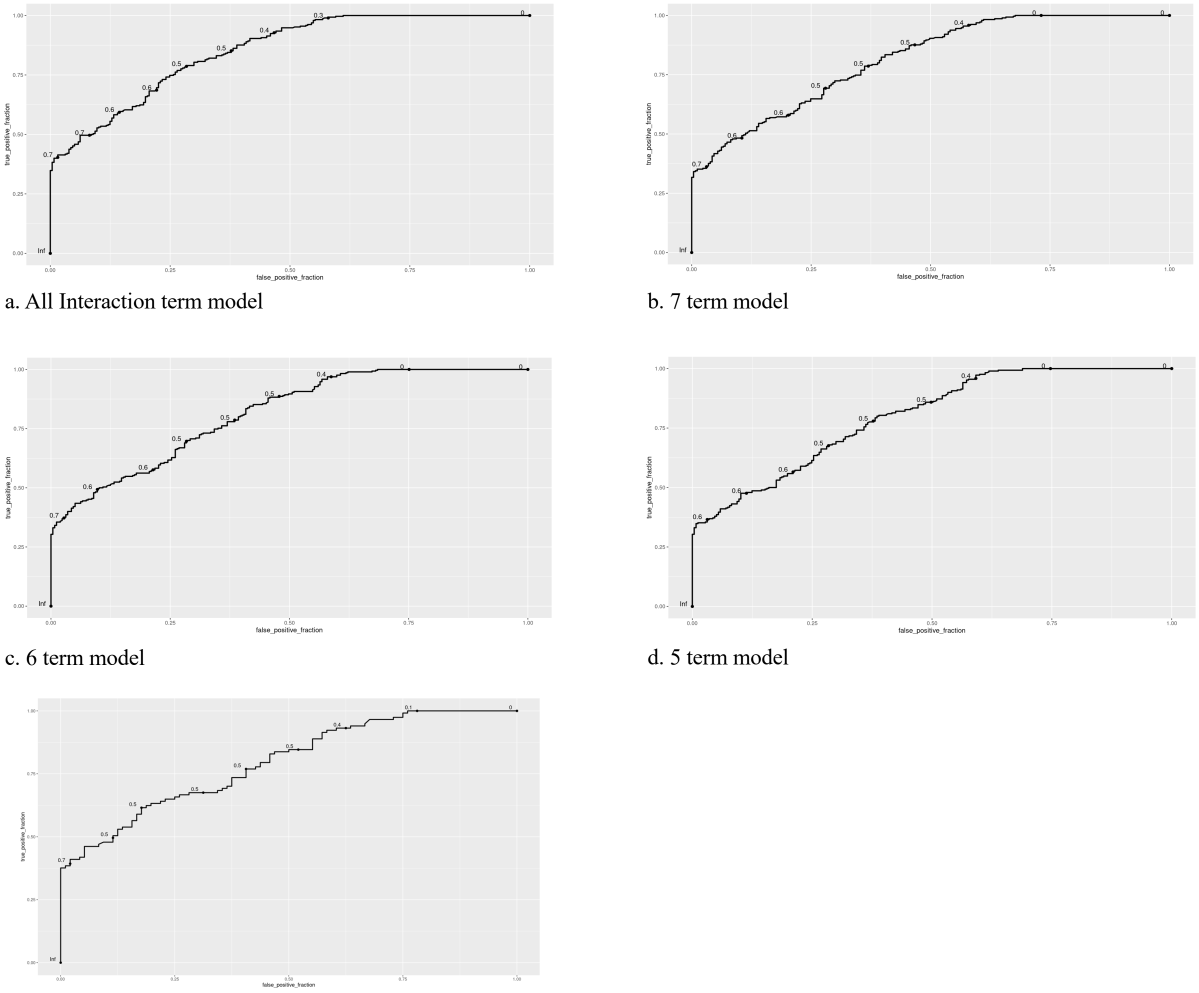
Wagner, C. E. van. (1987). *Development and structure of the Canadian forest fire weather index system*. Ottawa:Canadian Forestry Service.

Varela, N., Ospino, A., & Zelaya, N. A. L. (2020). Wireless sensor network for forest fire detection. *Procedia Computer Science*, *175*, 435-440.

Appendix

Model	AIC	Deviance	Accuracy	False Negatives
No-interaction Term Model	634.03	612.03	64.72%	28
All-interaction Term Model	573.61	463.61	74.41%	16

Appendix 1. Evaluation metrics for the no-interaction term model and the all-interaction term logistic regression models.



a. All Interaction term model

b. 7 term model

c. 6 term model

d. 5 term model

e. 5 term model (test)

Appendix 2. ROC curve for each model. Threshold of t=0.4 leads to between 90–95% sensitivity, with a false positive rate between 40–60%.

Full Model	Reduced Model	P-value
7 term model	6 term model	p = 0.1492
6 term model	5 term model	p = 0.0607
All-interaction term model	5 term model	p = 0.2544

Appendix 3. Drop-in-deviance test results.

Model	AIC	Deviance	Accuracy	False Negatives
5 terms	535.52	515.52	70.57%	7
All -interaction Term Model	573.61	463.608	74.41%	16

Appendix 4. Evaluation metrics for the 5 term and all-interaction term logistic regression models.

		Predicted Outcome	
Actual Outcome		No Fire	Fire
	No Fire	103	154
	Fire	7	283

Appendix 5. Confusion matrix for 5 term model.

		Predicted Outcome	
Actual Outcome		No Fire	Fire
	No Fire	108	149
	Fire	11	279

Appendix 6. Confusion matrix for 6 term model.

		Predicted Outcome	
Actual Outcome		No Fire	Fire
	No Fire	111	146
	Fire	15	275

Appendix 7. Confusion matrix for 7 term model.

		Predicted Outcome	
Actual Outcome		No Fire	Fire
	No Fire	133	124
	Fire	16	274

Appendix 8. Confusion matrix for all-interaction term model.

		Predicted Outcome	
Actual Outcome		No Fire	Fire
	No Fire	25	71
	Fire	3	114

Appendix 9. Confusion matrix for 5 term model on testing data.

Variables	**5 term model**	**6 term model**	**7 term model**	**All-interaction term model**
Temperature	Yes	Yes	Yes	Yes
Humidity	Yes	Yes	Yes	Yes
Windspeed	No	Yes	Yes	Yes
Rain	No	No	No	Yes
F.Moisture	Yes	Yes	Yes	Yes
D.Moisture	No	No	No	Yes
Drought	No	Yes	Yes	Yes
Initial.Spread	Yes	Yes	Yes	Yes
Region	Yes	Yes	Yes	Yes

Appendix 10. Terms included in each model.

Model	**Terms**
All-interaction term	(Temperature, Humidity, Windspeed, Rain, F.Moisture, D.Moisture, Drought, Initial.Spread, Portugal, SidiBelAbbes, Bejaia)^2
No-interaction term	Temperature, Humidity, Windspeed, Rain, F.Moisture, D.Moisture, Drought, Initial.Spread, Portugal, SidiBelAbbes, Bejaia
5 term	Temperature temperature*humidity temperature*F.moisture temperature*Initial Spread Initial Spread*Portugal
6 term	Temperature temperature*humidity temperature*F.moisture temperature*InitialSpread InitialSpread*Portugal windspeed*drought

7 term	Temperature temperature*humidity temperature*F.moisture temperature*InitialSpread Initial Spread*Portugal windspeed*drought humidity*F.Moisture

Appendix 11. Terms listed by each model.

Terms in 5 term model	**Estimate**	**Std. Error**	**Z value**	**p-value**
(Intercept)	-32.119068	18.014557	-1.783	0.0746
Temperature	-0.303441	0.359633	-0.844	0.3988
Humidity	-0.018956	0.020028	-0.947	0.3439
F.Moisture	0.023880	0.059441	0.402	0.6879
Initial.Spread	12.121778	6.518884	1.859	0.0630
Portugal	32.453103	17.191091	1.888	0.0591
Temperature:Humidity	0.001347	0.001159	1.162	0.2451
Temperature:F.Moisture	0.002015	0.003977	0.507	0.6123
Temperature:Initial.Spread	0.010888	0.005538	1.966	0.0493
Initial.Spread:Portugal	-12.378363	6.519025	-1.899	0.0576

Appendix 12. Coefficients for all terms in the 5 term model along with their standard errors, Z, and p-values.