Enhancing Forest Fire Prediction Using Machine Learning Techniques

Yuxin Liu (T00733525)
Qiuhan Li (T00728225)

Erfan Hoque
DASC 5420 Theoretical Machine Learning

2024-04-11

Abstract

This study examines the efficacy of various machine learning models in predicting the area affected by forest fires. This study uses a dataset comprising observations from Portugal's Montesinho Park. A comprehensive analysis involving Multiple Linear Regression, Stepwise Regression, Lasso, Ridge, Elastic Net, and Random Forest models was conducted. Initial analysis included variable transformation and multicollinearity assessment through VIF calculations and correlograms. Random Forest presents as the best model with the lowest RMSE, which indicates its potential as a reliable tool for developing forest fire predictions. The findings underscore the significance of model selection in forecasting burned areas. Further research may explore additional methods like SVM and neural networks and expand the scope to different geographical regions.

1. Introduction

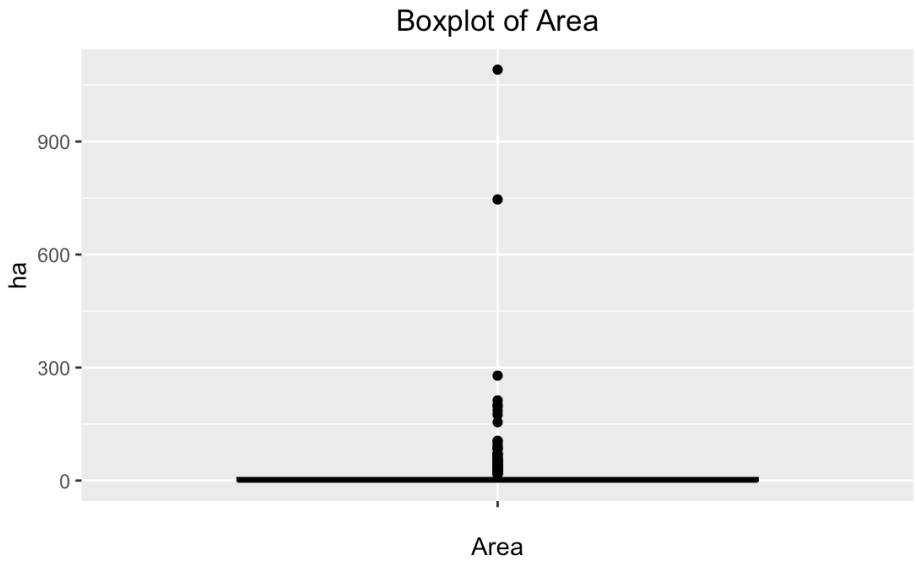
Forests are precious natural resources which are indispensable to humans. Forest fires pose a severe risk to natural landscapes and human habitats. Therefore, the prediction of wildfires is critically important for early warning systems and mitigation strategies. This project explores various statistical and machine learning methods to forecast the area affected by forest fires. We use environmental and weather-related predictors available in the forest fire dataset in UCI [1]. Our motivation is to determine which model best predicts fire extent, which can contribute to enhanced fire management and response initiatives.

2. Data

The dataset [1] includes observations of forest fires in the Montesinho Park in Portugal. It has variables such as x, y-axis spatial coordinates, month, day, the Forest Fire Weather Index (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, relative humidity (RH), wind speed, rain, and the burned area of the forest. We chose the variable, area, as our response variable. Other variables are considered as features.

2.1 Variable transformation

Firstly, we extracted the response variable, area, from the dataset. Next, we calculated the skewness and kurtosis to assess the data's distribution characteristics. The analysis helps us understand the degree of asymmetry observed in the distribution and its tailedness. To visually present these characteristics, we created a boxplot of the area variable (Plot 1). Based on the plot, we observed that most values are concentrated around 0, with few values sparsely distributed as individual points at higher values.



Plot 1 Boxplot of Area

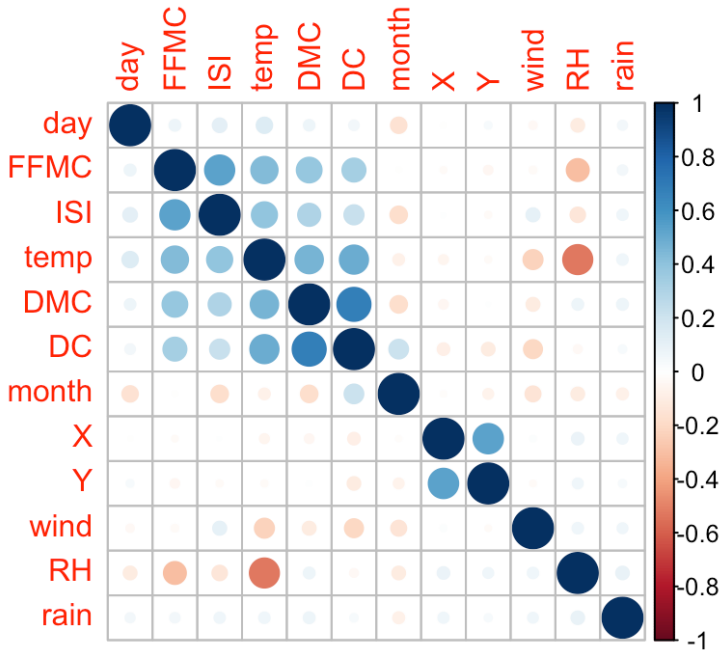
Based on our results, the area variable's distribution is highly skewed to the right, with most of the data concentrated in zero. This type of distribution may affect the performance of statistical and machine-learning models. Therefore, the area variable needs to be transformed into a more normal distribution for further model selections and prediction analysis.

We chose log translation, which can improve the effectiveness and stability of the model's predictions. Before using a log translation, we added 1 to the original values to avoid failing log translation. Then, we used the "ln" formula to transfer the area variable. This transformation is commonly used to handle data with skewed distributions to make them more normally distributed.

2.2 Multicollinearity

For further analysis, we need to check if there is multicollinearity between features. We plotted a correlogram and calculated the variance inflation factors (VIF) to identify the multicollinearity results.

The correlogram (Plot 2) illustrates the interrelations among different variables pertinent to forest fire data. Notably, there is a positive correlation between the Duff Moisture Code (DMC) and the Drought Code (DC). It indicates that as the moisture content measured by DMC decreases, the drought conditions reflected by DC intensify, which may jointly impact fire severity. Additionally, there is a negative correlation between temperature and relative humidity (RH). It shows that higher temperatures will decrease area's relative humidity, which makes sense. This visual correlation analysis is instrumental for identifying key predictors to be considered in the development of predictive models for forest fire susceptibility and behavior.



Plot 2 Graph of the Correlation Matrix

The Variance Inflation Factor (VIF) results (Table 1) suggest that multicollinearity is present among the predictors within the forest fire dataset. Specifically, the Duff Moisture Code (DMC) exhibits the highest VIF value at approximately 2.709, which indicates a relatively high correlation with other variables and potential redundancy. The Drought Code (DC) and temperature (temp) have VIF values of around 2.652 and 2.708, respectively. These values exceed the common threshold of 2.5, which suggests that these variables may be contributing to multicollinearity in the model. However, because the VIF values are all smaller than 5, so we will not move forward with analyzing the multicollinearity of these features.

Table 1 VIF Summary Table

Feature	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain
VIF	1.443	1.438	1.465	1.063	1.683	2.709	2.652	1.622	2.708	1.922	1.162	1.064

3. Method

The methodologies focus on predicting the area affected by forest fires. Our approach involves various statistical and machine learning models and validation techniques to evaluate model performance. (The code on GitHub: https://github.com/YuxinLiu-Adeline/DASC-5420-Final-Project)

3.1 Model Implementation

We predicted the area of forest fires with several models, with each model implemented as such:

Multiple Linear Regression: The multiple linear regression (MLR) includes all features and serves as the baseline model for comparison. It was employed to identify how various predictors, such as weather conditions and topographical features, influence the area of forest fires. This model aims to fit a linear equation to these factors to predict the fire's extent.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

This formula represents a linear equation where Y is the dependent variable predicted by the intercept β_0, the predictor variables $X_1, X_2, \ldots, X_n$, their respective coefficients $\beta_1, \beta_2, \ldots, \beta_n$, and an error term ε.

Stepwise Regression: We used the backward elimination method with the Akaike Information Criterion (AIC) as a stopping rule. This approach helped us identify the most significant predictors by removing predictors based on their statistical significance and impact on model performance.

Start with full model:
Iteratively remove (backward) variables:
Backward Elimination: $Y = \beta_0 + \sum_{i \neq j} \beta_i X_i + \ldots + \varepsilon$

Here, $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients to be determined for predictor variables $X_1, X_2, \ldots, X_p$, and ε is the error term. The stepwise process decides which β_i coefficients remain in the model based on AIC.

Lasso Regression: Lasso Regression, which stands for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that uses shrinkage. Lasso can both select important features of the dataset and shrink the coefficients of less important ones to zero, effectively removing them from the equation. The strength of the penalty applied to the features is adjustable via a tuning parameter, λ. We used 10-fold cross-validation to determine the best λ.

Ridge Regression: Similar to LASSO, Ridge Regression is used for feature shrinkage and to combat multicollinearity by penalizing the size of the regression coefficients. Ridge uses the square of the coefficients β in the penalty term, which is controlled by a tuning parameter λ. The Ridge Regression does not enforce sparsity in the model. Hence, it will not set coefficients to zero but will shrink them close to zero. The optimal value of λ will be found by using 10-fold cross-validation. The optimal value aims to find a balance between bias and variance.

Elastic Net Regression: Elastic Net Regression combines the properties of LASSO and Ridge regression methods by penalizing the model with both $L1$ and $L2$ regularization terms. The technique is capable of feature selection like LASSO while still stabilizing the model coefficients like Ridge when predictors are correlated. In Elastic Net, two parameters control the model complexity. The parameter λ manages the overall strength of the penalty. And the parameter α balances the proportion of $L1$ versus $L2$ penalty. Optimal values for these parameters are typically selected through a cross-validation process, aiming to minimize prediction error and prevent overfitting. We use 10-fold cross validation with different α values to find optimal values.

Random Forest: Random Forest is an ensemble technique that constructs numerous decision trees to improve predictive accuracy and mitigate overfitting. It introduces randomness during tree generation and aggregates individual trees' predictions to produce an outcome. The random forest model can efficiently handle large datasets with complex structures by optimizing key parameters – such as the number of trees and the subset of features used for splitting – through

cross-validation. The article “Prediction and driving factors of forest fire occurrence in Jilin Province, China” [2] emphasized the random forest method's high accuracy in predicting forest fire data. We grew 500 trees, randomly sampling 3 variables as candidates at each split.

3.2 Model Validation and Performance Evaluation

To validate the models and evaluate their performance, the following methods were used:

Cross-Validation: We employed 10-fold cross-validation for our models. This technique involves partitioning the data into ten subsets and holding out each subset in turn as a test set while training on the remaining nine. The process is repeated ten times, with each subset used once as the test set. This method helps us to understand the models' stability and performance across different subsets of data.

Performance Metrics: The root mean squared error (RMSE) was used to evaluate the models. This is a measure of the average magnitude of the model's validation errors. A lower RMSE means a better model's predictive accuracy. Models were evaluated based on their root mean squared errors from predictions in testing data compared to actual observed values.

4. Results

Multiple linear regression (MLR) has a validation RMSE of 1.855, which was assigned as the baseline value. After using backward stepwise regression to optimize for the lowest Akaike Information Criterion (AIC), we identified a more parsimonious model featuring Duff Moisture Code (DMC) and Initial Spread Index (ISI) as significant predictors. This stepwise model is better than MLR with a lower validation RMSE of 1.410.

Table 2 Summary Table of RMSE

Method	RMSE
Multiple Linear Regression	1.855
Stepwise Regression	1.410
LASSO Regression	1.407
Ridge Regression	1.420
Elastic Net Regression	1.407
Random Forest	1.371

The regularization techniques, including LASSO, elastic net, and Ridge regression, delivered comparable results with RMSE values around 1.41 and 1.42. These outcomes exhibited similar improvement as stepwise regression did.

The Random Forest model outperformed all the models with the lowest RMSE of 1.371. This indicates its high accuracy in predicting forest fire sizes and validates its efficacy in handling complex datasets.

5. Discussion and Conclusion

In conclusion, stepwise regression improved upon this with a more selective model featuring Duff Moisture Code and Initial Spread Index variable. Stepwise regression has similar results as regularization methods such as LASSO, elastic net, and Ridge regression, which is around an

RMSE of 1.41. Compared with multiple linear regression, their outcomes enhance model performances.

The Random Forest model is the best model, with the lowest RMSE of 1.371. This result emphasizes its higher predictive accuracy. Compared with other predictive models, it confirms the strength of the Random Forest algorithm in capturing complex patterns and interactions among various predictors. Hence, we can conclude that random forest is a reliable tool in the development of forest fire predictions.

For further studies, researchers can, based on the results, explore more methods like support vector machine (SVM) and neural networks. Additionally, further studies can explore models for forest fire data in different countries.

References

1. Cortez P, Morais A. Forest fires [Internet]. 2008 [cited 2024 Apr 17]. Available from: https://archive.ics.uci.edu/dataset/162/forest+fires

2. Gao B, Shan Y, Liu X, Yin S, Yu B, Cui C, et al. Prediction and driving factors of forest fire occurrence in Jilin province, China. Journal of Forestry Research. 2023 Dec 16;35(1). doi:10.1007/s11676-023-01663-w