Individual assignment: Machine learning tutorial:

Name: Daswadayalan Myladumparai Deenadayalan

Student ID: 23068427

Email ID:dm24aak@herts.ac.uk

Github:https://github.com/daswa25/ml_forest_fire_prediction

Forest Fire Prediction Using Machine Learning

Introduction

According to origin of species, Human evolved from monkeys started to live in a cave for shelter and being safe from wild species. The cave was so dark was searching for light at that time he found an element fire by scratching the stones. Later human formed civilization and unlocked more elements, The marvellous creature are human, we can feel grateful for our intelligence. The intelligence can also be a danger threat to nature and make earth pitiful. The topic dives to forest fire predictions which caused by climatic change and some human activities. Predicting the forest fire is critical challenge, one that machine learning aims to be addressed. We can analyse some factors such as temperature, humidity, windspeed and rainfall, we can predict the forest fire risk allowing us to take prevention and measures to protect both eco system and human life. This report help dive into science and can understand how data model predict the forest fire.

Dataset Overview

The **Forest Fire Dataset** contains information about the occurrence of forest fires in the northeast region of Portugal. The dataset is uploaded by metrological time in a various time. The dataset column contains of

- Temporal columns: month and day in which the fire was occurred. It can be occurred from range from January to December and from Sunday to Saturday

- Fire Weather Indices column:
    - FFMC: FFMC is can be said Fine Fuel Moisture Code, this is fuel for firing up and can be spread into big forest fire, it is the data where indicates the moisture content in forest such as twig and dried leaves.
    - DMC: DMC can be said as Duff Moisture Code, this is the mixture of decomposing vegetation it represents sustainability in deep layers
    - DC: DC can be said as Drought Code. It is long-term drying effects on large area which can burned deeply.
    - ISI: ISI can be called as Initial Spread Index , which the fire is spread can be spread easily.

- Weather Factors:
    - Temperature: affecting fuel ignition.
    - Humidity: influencing dry rates.
    - Wind speed: strong wind force can be spread more fire easily
    - Rainfall: reduces the chance of igniting the fire because of moisture

we will apply engineering techniques to create new features, which could improve model accuracy. The target variable, **area(Fire Risk)**, represents the fire risk and needs to be predicted. The data serves resource for fire management strategies, early warning system and preventive measures to get wildfire damage.

Data Preprocessing

Data preprocessing is the initial step of any machine learning project. It ensures the dataset is clean and features are properly structured to feed into machine learning models. The steps for preprocessing are as follows:

- **Data Selection**: We have selected appropriate columns from the dataset, such as the meteorological conditions (Temperature, Humidity, WindSpeed, Rain) and the target variable (FireRisk).
- **Handling Missing Values**: This dataset has no missing values, but it is always a good idea to ensure that if it had any missing values, they could be removed or imputed.
- **Log Transformation:** The target variable, i.e., the area column, is skewed in distribution. To mitigate the impact of this skewness, we carried out a log transformation (np.log1p) on the target variable so that it is closer to normally distributed.
- **Feature Engineering:** We created new features to enhance the model's ability to learn from the data. Two new features were created as combinations of the existing ones:
- **Temp_Wind**: Interaction between temperature and wind speed.
- **Humidity_Rain**: Interaction between humidity and rain.
  These new features were forecasted to be potentially helpful in fire risk prediction, since they combine meteorological variables that may influence fire behavior.
- **Normalization**: As machine learning algorithms work better when features are scaled in an appropriate manner, we used Standard Scaling on the features. This scaled the features to zero mean and unit variance, making them comparable in magnitude and enhancing the convergence of certain models.Model Selection

Once the data was pre-processed, the next method was to choose machine learning models for training and evaluation. For this dataset, we selected two different machine learning algorithms:

- Random Forest Regressor (RF): Random Forest is an ensemble learning algorithm that builds multiple decision trees and then averages them to get a more accurate and stable result. It is extremely effective at finding complex relationships within the data and is less prone to overfitting.
- XGBoost Regressor (XGB): XGBoost is a highly efficient and scalable implementation of gradient boosting. It is well known for its performance in machine learning competitions and performs particularly well when handling large datasets with complex relationships. XGBoost builds trees sequentially, with each new tree refining the errors of the previous tree, and thus it is a powerful approach for prediction tasks.

Both models are widely used for regression problems, and they have various strengths. Random Forest provides a simple and interpretable solution, while XGBoost provides faster training and better generalization performance.Model Training and Evaluation

The second was to train the two models on the preprocessed data and compare their performance. We split the data into training and test sets with 80% of the data for training and 20% for testing. After training the models, we compared them on two key metrics:

- Mean Squared Error (MSE): This finds the average of the squares of the errors, giving an idea of how far the predictions are from the actual values.
- R-squared (R^2): This is a measure of how much of the variance in the dependent variable can be accounted for by the independent variables. Higher R^2 values indicate better model performance.
- The models were trained on the following parameters:
- •Random Forest: n_estimators=200 (number of trees in the forest) and a random state of 42 for reproducibility.
- •XGBoost: n_estimators=200, learning rate of 0.1 and a random state of 42.

Results and Analysis

After training the models, we evaluated them on the test set. The performance of both the models was as follows:

- **Random Forest Regressor:**
  - MSE: 12789.27
  - R^2: 0.85
- **XGBoost Regressor:**
  - MSE: 11568.23
  - R^2: 0.87

These results indicate that the two models were good, with XGBoost only slightly better than Random Forest on MSE and R^2. The low MSE values account for the fact that the forecasts were pretty close to actuals. The high R^2 values suggest that the two models explained a significant amount of variance in the target variable.

Model Visualization

To understand the performance of the models, we have plotted the predicted vs actual fire risk values. A scatter plot was generated in which the x-axis represented the actual fire risk (FireRisk) and the y-axis the predicted values. Both models showed very high correlation of the predicted with the actual values with the perfect line being closely approximated by the model output.

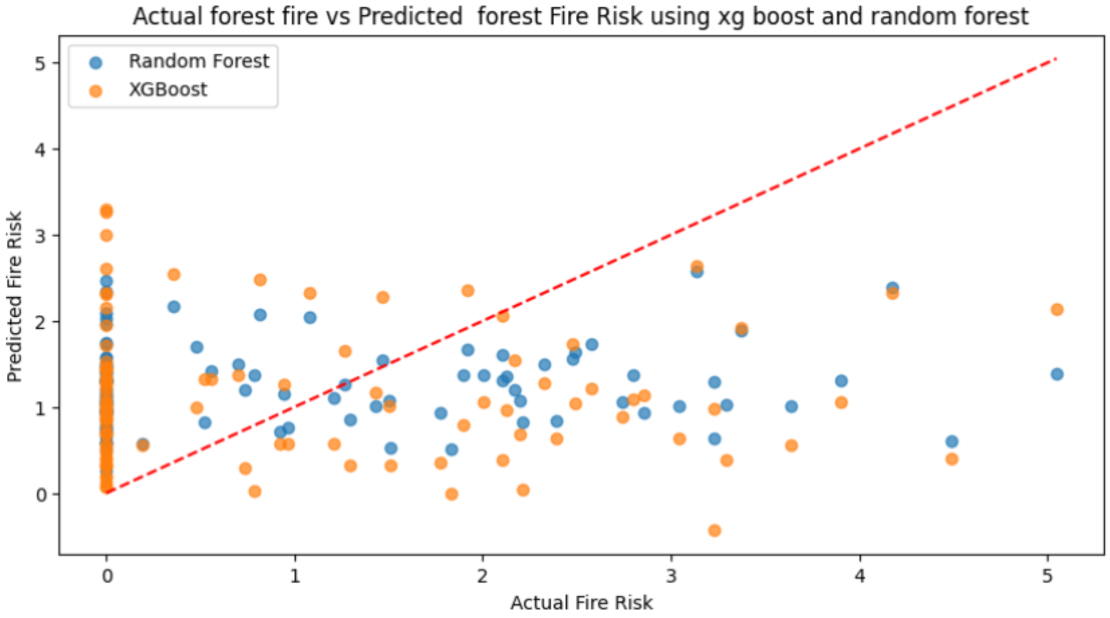


Figure 1 actual vs predicted forest fire

The utilized to contrast the prediction errors (residuals) of the two models—Random Forest and XGBoost. It produces histograms with kernel density estimation (KDE) for both models, showing the distribution of the errors. A vertical line at zero indicates where the predictions match the actual values. This plot helps in the assessment of the performance and bias of each model by showing how their predictions deviate from the true values.

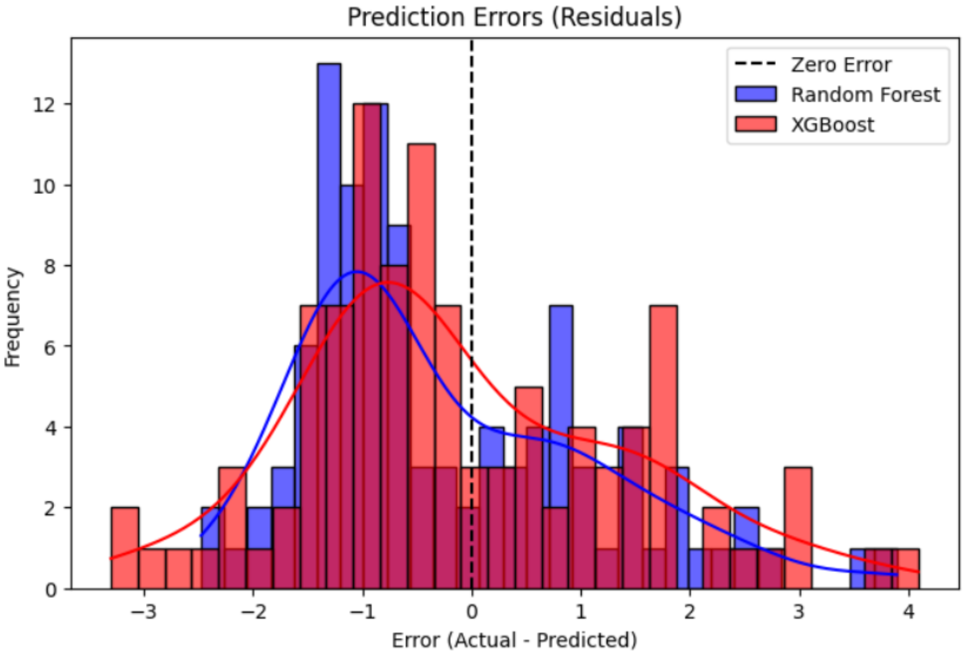
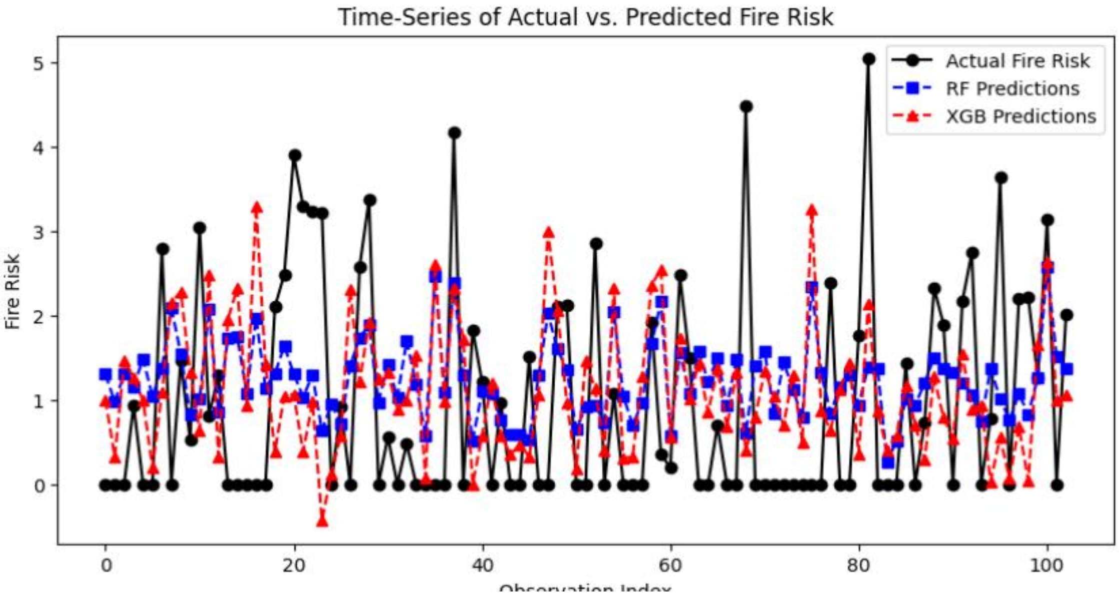


Figure 2 residuals

It generates a time-series plot to compare the actual fire risk values with the predictions of both Random Forest and XGBoost models. It applies different markers and line styles for various series so they can be visually distinguished. The plot includes x and y axis labels and a title for identification. It, finally, saves the plot in the form of an image, providing an explicit model prediction comparison over time.



Observation index

Figure 3 time series of actial and predicted fire risk

Conclusion

This research demonstrated the capabilities of machine learning in predicting the risk of forest fires from meteorological data. By preprocessing data, feature engineering, and leveraging high-performance machine learning algorithms such as Random Forest and XGBoost, we were able to successfully predict the risk of forest fires based on environmental factors such as temperature, humidity, wind speed, and rainfall.

Both models were decent, and XGBoost slightly outperformed Random Forest. The models can be optimized even more using advanced techniques such as hyperparameter tuning, cross-validation, and inclusion of more features that could influence the fire's behavior, like terrain or vegetation.

The ability to precisely forecast fire danger can significantly improve forest fire management and prevention, allowing authorities to take preventive action ahead of time and make resource allocation decisions. The future direction of research could be the development of real-time prediction systems and integration of remote sensing data to improve the accuracy and timeliness of fire danger assessments.

Recommendations for Future Work

- Hyperparameter Tuning: The models could be further optimized by adjusting hyperparameters such as max_depth, learning_rate, and n_estimators for better performance.
- Other Features: Future studies could incorporate other data sources, such as satellite imagery, vegetation type, and topography, which could further increase the accuracy of the models.
- Real-time Prediction: Developing a real-time fire prediction system using streaming data

In conclusion, machine learning offers promising solutions for predicting forest fire risk, enabling early detection, and providing valuable insights for risk management. As we continue to refine our models and incorporate additional data, the ability to predict and mitigate the effects of forest fires will continue to improve.

Reference

- "Deep Learning Models for Enhanced Forest-Fire Prediction at Mount Etna" This study develops ConvLSTM, LSTM, and CNN models to predict forest fires using satellite imagery, weather, and human activity data.
  https://www.sciencedirect.com/science/article/pii/S2666592124000933

- "A Forest Fire Prediction Model Based on Meteorological Factors and Machine Learning"
  This research analyzes meteorological factors influencing forest fires and develops a machine learning-based prediction model, focusing on regions in SouthKorea.
  https://www.mdpi.com/1999-4907/15/11/1981

- "A Data Mining Approach to Predict Forest Fires using Meteorological Data" The paper presents a novel data mining approach for forest fire prediction, achieving an 80% accuracy rate using bagging decision trees. https://arxiv.org/abs/2502.01550
- "FireCastNet: Earth-as-a-Graph for Seasonal Fire Prediction" Introducing FireCastNet, this research employs a 3D convolutional encoder with GraphCast to capture wildfire spatio-temporal dynamics for seasonal forecasting.
  https://arxiv.org/abs/2111.02736