Predicting Forest Fire Occurrence Using LightGBM

Introduction

Forest fires pose a significant threat to ecosystems, economies, and human lives globally. Accurate prediction of forest fire occurrences can aid in effective resource allocation and preventive measures. Machine learning techniques have been increasingly applied to environmental data to predict such natural disasters. In this project, I aim to develop a predictive model using Light Gradient Boosting Machine (LightGBM) to classify the likelihood of forest fires based on meteorological and environmental features. I utilize the publicly available Forest Fires Dataset from the UCI Machine Learning Repository (Cortez & Morais, 2007).

Method

Dataset

The Forest Fires Dataset contains 517 instances with the following attributes:

- Spatial data: X and Y axis coordinates.
- Temporal data: month and day.
- Meteorological data: FFMC, DMC, DC, ISI (various fire weather indices), temperature, relative humidity, wind, and rain.
- Target variable: area burned by the fire (in hectares).

For this project, I transformed the area variable into a binary classification target:

- 0: No fire (area burned is 0).
- 1: Fire occurred (area burned > 0).

Data Preprocessing

1. Handling Outliers: Applied the Interquartile Range (IQR) method to cap outliers in numerical features.
2. Feature Scaling: Standardized numerical features using StandardScaler to normalize their ranges.
3. Encoding Categorical Variables: Encoded month and day using Label Encoding to convert categorical data into numerical format.
4. Feature Engineering:
    - Created temp_rain_ratio to capture the interaction between temperature and rain.
    - Created wind_rain_interaction to represent the combined effect of wind and rain.

Algorithm Justification

I selected LightGBM for its efficiency and high performance with tabular data, especially for classification tasks with imbalanced datasets. LightGBM handles categorical features and large datasets effectively and supports parallel and GPU learning.

Model Training

- Data Splitting: Split the dataset into training (80%) and testing (20%) sets with stratification to maintain class balance.
- Hyperparameter Tuning: Performed Grid Search with 5-fold cross-validation to optimize hyperparameters:
    - num_leaves: [20, 50, 100, 200, 400]
    - max_depth: [5, 8, -1]
    - learning_rate: [0.005, 0.01, 0.05, 0.1]
    - n_estimators: [20, 50, 100]
    - feature_fraction: [0.2, 0.6, 0.8]

Best Parameters Found

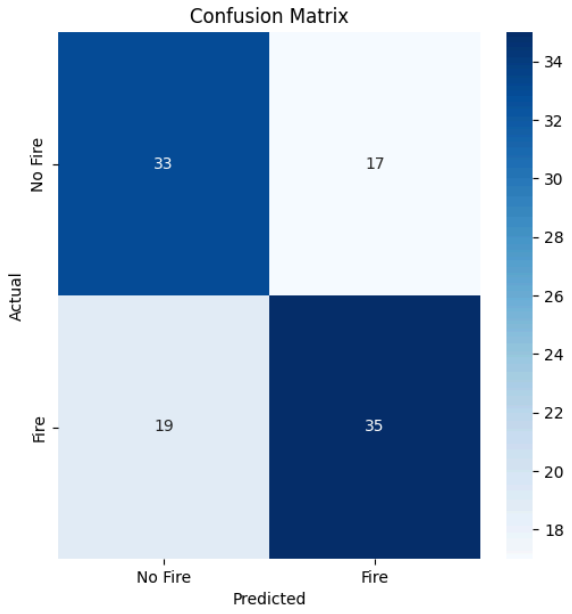
feature_fraction: 0.8,
learning_rate: 0.05,
max_depth: 8,
n_estimators: 50,
num_leaves: 20

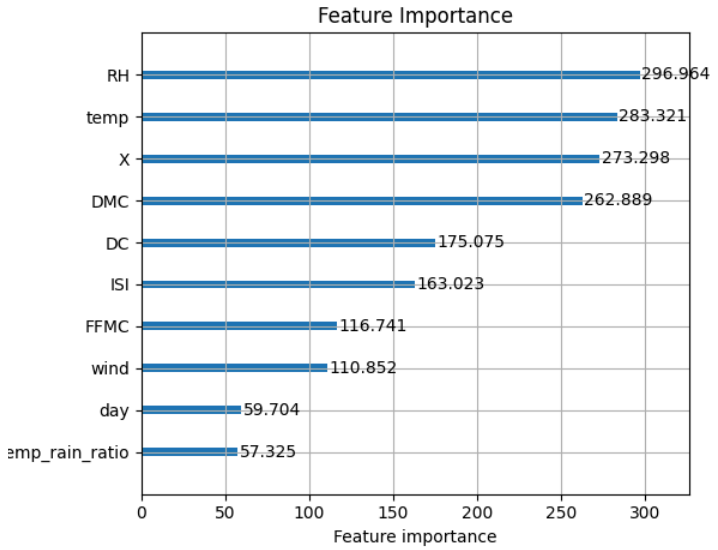
Results

Evaluation Metrics

- Accuracy: 0.6538
- Precision: 0.6731
- Recall: 0.6481
- F1 Score: 0.6604
- ROC AUC: 0.6513

Confusion Matrix and Feature Importance





Discussion

The model achieved moderate performance with 65% accuracy and a ROC AUC of 0.6513, showing fair separability between fire and no-fire classes. Precision and recall indicate the model is slightly better at predicting fire occurrences than their absence.

Feature Importance:

1. RH (Relative Humidity) and Temperature were the most significant features, highlighting the role of meteorological conditions.
2. X (spatial coordinate) suggested the importance of geographic influences.
3. DMC and DC, key fire weather indices, validated their correlation with fire behavior.

Model Limitations:

- Small Dataset: The model's ability to generalize is limited by the small data size.
- Geographic Specificity: The model may only perform well for the specific location represented in the dataset, reducing its utility for broader applications.
- Moderate ROC AUC: Indicates the need for further optimization.

Potential Impacts:

- Positive:
    - Assists in early detection and prevention of forest fires.
    - Helps allocate resources efficiently for fire management.
- Negative:
    - Over-reliance on the model without considering other environmental factors could lead to misallocation.

Future Work:

- Incorporate additional relevant features (e.g., vegetation type, human activity data).
- Experiment with more advanced models or ensemble methods.

References

- Cortez, P. & Morais, A. (2007). Forest Fires [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5D88D.
- Cortez, P., & Morais, A. (2007). *A Data Mining Approach to Predict Forest Fires using Meteorological Data*. In *Proceedings of the 13th Portuguese Conference on Artificial Intelligence* (EPIA 2007), December, Guimaraes, Portugal (pp. 512-523). DOI: 10.1007/978-3-540-77226-2_39.