Advancements in Forest Fire Prediction
Integrating AI and Statistical Inference

Presented By
Mounika Malka

Introduction

C

- Forest fires pose significant threats to ecosystems, lives, and infrastructure. The increasing frequency and severity of forest fires underscore the need for effective prediction methods. Conventional prediction systems are costly and inaccurate, especially in developing countries. Efficient and cost-effective prediction methods are essential for mitigating risks.

C

• Predicting forest fires is complex due to various environmental factors. Conventional systems rely heavily on extensive monitoring, leading to inaccuracies. Inaccurate weather predictions contribute to errors in fire risk assessment. Developing reliable prediction methods poses challenges, especially for developing countries.

Q

•How can machine learning techniques enhance forest fire prediction accuracy?
•What methodologies can address the challenges in developing reliable prediction systems?
•How do reduced parameter sets and statistical inference techniques improve prediction efficiency?
•What role do cost-effective methods play in mitigating forest fire risks, particularly in developing countries?

H

•H: Machine learning techniques will significantly improve forest fire prediction accuracy compared to conventional methods. The integration of reduced parameter sets and statistical inference techniques will enhance the efficiency of forest fire prediction models. Cost-effective prediction methods will be developed, addressing the challenges faced by developing countries. Comprehensive methodologies will lead to the development of reliable prediction systems for forest fire management

Questions….?

Why this research is important	This research is crucial for mitigating the devastating effects of forest fires and safeguarding both human and environmental well-being.
What we know and don't know	While Past data patterns, Model performance and Preprocessing techniques provide valuable insights, there are still uncertainties regarding Future fire dynamics and The optimal approach to predicting and mitigating forest fires. Ongoing research is needed to address these knowledge gaps.
Experiment	Our experiment aimed to develop accurate regression models for forest fire prediction by leveraging machine learning techniques and analysing relevant meteorological data.
Hypothesis	We hypothesize that by utilizing advanced regression algorithms and comprehensive meteorological data, we can develop predictive models capable of accurately forecasting forest fire occurrences.

Design/Methods

Data Collection

A dataset was collected from the northeast region of Portugal, comprising meteorological variables such as temperature, humidity, wind speed, and precipitation, along with historical records of forest fire occurrences.

Model Training and Evaluation

The collected data was divided into training and testing sets for model training and evaluation.
Linear regression, Xgboost regressor, Catboost regressor and lightbgm regressor

Cross-validation techniques, such as k-fold cross-validation, were employed to assess the models' performance robustly.
Various metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared score (R2), were used to evaluate the models' performance.

Feature Importance Analysis

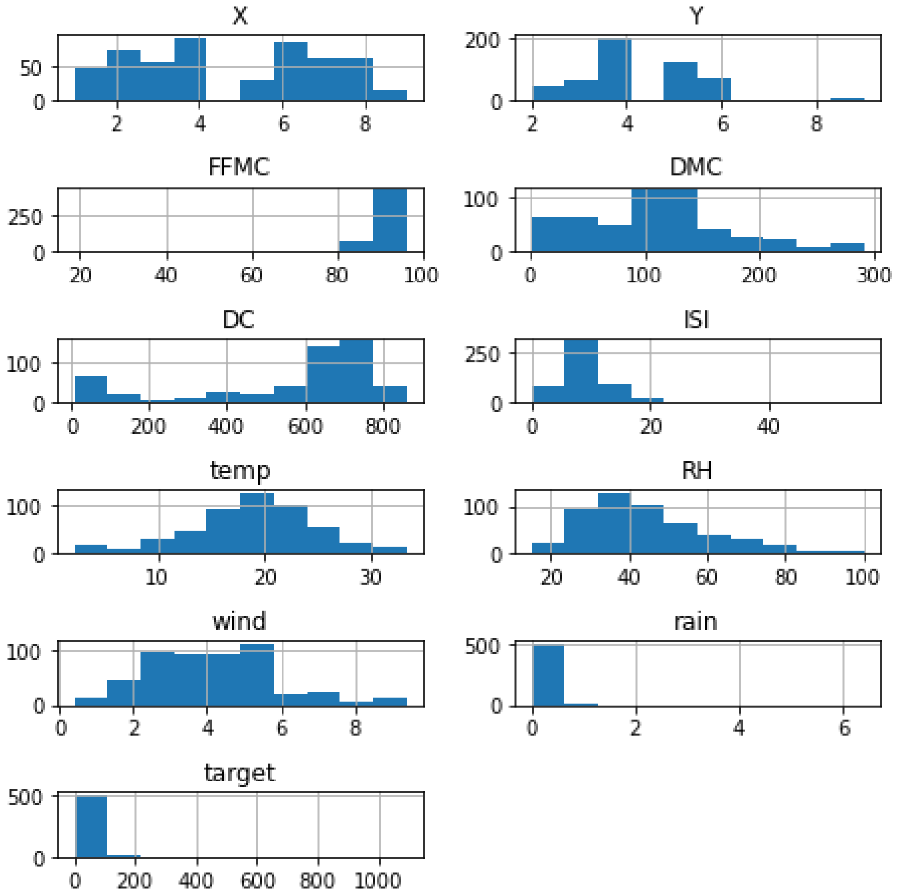
Feature importance analysis was conducted to identify the most influential features for predicting forest fire area.
Techniques such as permutation importance or feature importance scores provided by the models were used for this analysis.

Design/Methods

Comparison of Top Dependable Variables vs. All Variables:

The performance of the models using only the top 5 dependable columns versus using all columns was compared to assess the impact of feature selection on prediction accuracy.

Histograms for each column in the DataFrame using a specified layout. Each histogram represents the distribution of values for a particular feature or variable.



Data

- Correlation Analysis:
- Conducted correlation analysis to understand the relationship between variables.
- Positive correlation: temperature (temp) showed the strongest correlation with the target variable.
- Negative correlation: relative humidity (RH) exhibited the weakest correlation

Feature Engineering Techniques

One-Hot Encoding:

Categorical variables such as day and month transformed into numerical form using one-hot encoding.

Enabled integration of categorical variables into regression models.

Binary Encoding:

Applied binary encoding to the rain column to simplify representation.

Facilitated inclusion of the rain variable in regression models.

Outlier Removal:

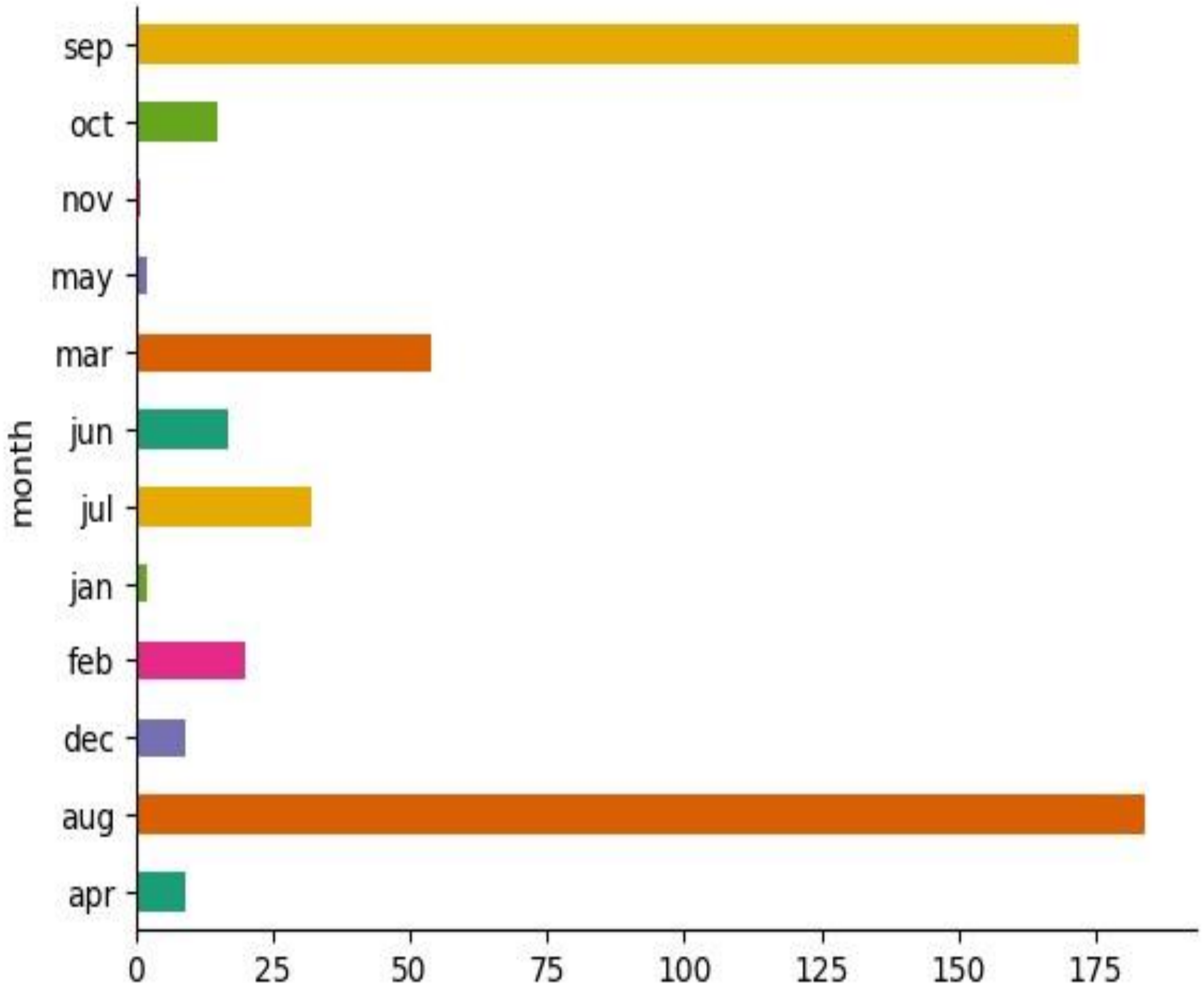
Identified and removed data points significantly deviating from the dataset's distribution.

Prevented outliers from skewing model predictions.

Log Transformation:

Conducted log transformation on the target variable (area) to achieve a more symmetric distribution.

Stabilized variance, a common practice in linear regression problems.



Data/Result

Model	Test size	Parameters	MSE	RMSE	MAE	R2Score	Features
linear regression	0.1		553.8311087	23.53361657	15.91335292	0.0562812748	
xgboost regressor	0.1		559.2097578	23.64761632	14.33117115	0.04711614884	
catboost regressor	0.1		599.7692371	24.49018655	14.70741921	-0.02199650927	
lightgbm regressor	0.1		919.2956132	30.31988808	20.03166733	-0.5664639824	
linear regression	0.2		1.887023901	1.373689885	1.148276851	0.006522653558	1) One hot encoding :- day ,month 2)binary:-rain column 3)remove outliers 4)log transform :-area
xgboost regressor	0.2	{'colsample_bytree': 0.6, 'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}	2.037411592	1.427379274	1.139923599	-0.07265321924	
catboost regressor	0.2	{'iterations': 100, 'learning_rate': 0.01, 'depth': 8, 'loss_function': 'RMSE'}	1.889531513	1.374602311	1.149396133	0.005202450237	
lightgbm regressor	0.2	{'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.05, 'max_depth': 4, 'min_child_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': None, 'num_leaves': 20, 'objective': None, 'random_state': None, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'subsample': 1.0, 'subsample_for_bin': 200000, 'subsample_freq': 0}	1.976040596	1.405717111	1.134576846	-0.04034271452	
random forest	0.1	max_depth: None min_samples_leaf: 4 min_samples_split: 10 n_estimators: 100	2.597105962	1.611553897	1.248812211	-0.06848152437	
random forest	0.2		1.976186474	1.405768997	1.142620107	-0.04041951613	
decision tree	0.1	max_depth: 10 min_samples_leaf: 4 min_samples_split: 10	3.360631143	1.833202428	1.365077308	-0.3826052303	
decision tree	0.2		2.919131019	1.708546464	1.241235005	-0.5368594628	
extra tree	0.1	max_depth: 10 min_samples_leaf: 4 min_samples_split: 10 n_estimators: 200	2.573769042	1.604297055	1.205008535	-0.05888042669	
extra tree	0.2		1.885591855	1.373168546	1.07551415	0.007276595008	

Result

Significance of the Findings:

- The results highlight the efficacy of machine learning in predicting forest fires when data preprocessing is done effectively.
- By leveraging historical weather, geographical, and fire data, these models uncover patterns to aid in fire prediction, enhancing preparedness and management.

Performance Analysis:

- CatBoost outperformed other models in predicting fires due to its adeptness in handling diverse data types and identifying complex patterns.
- XGBoost and LightGBM also performed well, but CatBoost exhibited slightly superior performance.
- Integration of an additional feature for predicting fire severity into three categories mild, moderate, and severe enhances the models' ability to anticipate fire intensity accurately, aiding in resource allocation and mitigation strategies.

Limitations and Implications:

- The analysis acknowledges limitations such as reliance on dataset quality and representativeness, emphasizing the impact of hyperparameters and preprocessing techniques on model performance.
- Challenges include data scarcity, biased predictions due to limited observations or missing values, necessitating innovative approaches like data imputation or supplementary data integration for addressing data deficiencies.

Conclusion

Effectiveness of Regression Models:

This study demonstrates the effectiveness of regression models in forest fire prediction.

Implementation of appropriate preprocessing techniques and model selection strategies significantly enhances predictive accuracy.

Performance Comparison:

CatBoost emerged as the top performer among the evaluated models.

Its robustness and adaptability make it a valuable tool for forest fire prediction and management.

Impressive metrics achieved by CatBoost include MSE=1.89, RMSE=1.37, MAE=1.15, and R^2=0.0052.

Future Work

- **Future research directions include:**
- Exploring advanced preprocessing techniques like feature engineering and dimensionality reduction.
- Integrating domain knowledge and external data sources such as satellite imagery and topographical information.
- Evaluating ensemble methods and hybrid models to enhance predictive performance.

Thank You!!!!