Flood Probability Prediction Analysis

1. Introduction

This project aims to predict flood probabilities using machine learning models. The analysis involves data preprocessing, feature engineering, model training, and evaluation. The datasets used include training and testing data with various features related to flood probability. The goal is to build robust models that can accurately predict flood probabilities, which can be crucial for disaster management and preparedness.

2. Data Loading and Initial Inspection

2.1. Import Libraries

I start by importing the necessary libraries for data manipulation, visualization, and machine learning. Libraries such as pandas, numpy, seaborn, matplotlib, and machine learning libraries like xgboost, catboost, and lightgbm are essential for this analysis.

2.2. Load Data

The training and testing datasets are loaded from CSV files. These datasets contain various features that are potential indicators of flood probability.

2.3. Inspect Data

I inspect the shape of the datasets to understand the number of rows and columns. The training dataset has 1,117,957 rows and 22 columns, while the testing dataset has 745,305 rows and 21 columns. I also print the first few rows of each dataset to get a sense of the data structure and content.

2.4. Check for Null Values

I check for null values in both datasets to ensure data integrity. Both datasets have no null values, which is a good sign for the quality of the data.

3. Data Preprocessing

3.1. Remove Outliers

Outliers can significantly affect the performance of machine learning models. I define a function to remove outliers using the Interquartile Range (IQR) method. This method identifies and removes data points that fall below the lower bound or above the upper bound, which are calculated based on the first and third quartiles. After removing outliers, the shape of the cleaned training dataset is reduced to 849,100 rows and 20 columns.

3.2. Feature Engineering

Feature engineering involves creating new features from the existing ones to improve the performance of machine learning models. I define a function to calculate statistical measures such as mean, standard deviation, maximum, minimum, range, variance, skewness, kurtosis, and median absolute deviation for each row. These new features can capture additional patterns and relationships in the data.

3.3. Data Scaling

Data scaling is essential for machine learning models that are sensitive to the scale of the input features. I use the StandardScaler from sklearn to standardize the features by removing the mean and scaling to unit variance. This ensures that all features contribute equally to the distance calculations in the models.

4. Model Training and Evaluation

4.1. CatBoost Regressor

CatBoost is a gradient boosting library that is efficient and provides state-of-the-art results. I split the data into training and testing sets, train the CatBoost model, and evaluate its performance using the R-squared metric. The R-squared value for the CatBoost model is approximately 0.9267, indicating a high level of accuracy in predicting flood probabilities.

4.2. LightGBM Regressor

LightGBM is another gradient boosting framework that is designed to be distributed and efficient. I train the LightGBM model and evaluate its performance using the R-squared metric. The R-squared value for the LightGBM model is also high, indicating its effectiveness in predicting flood probabilities.

4.3. XGBoost Regressor

XGBoost is a popular gradient boosting library that is known for its performance and speed. I train the XGBoost model and evaluate its performance using the R-squared metric. The R-squared value for the XGBoost model is approximately 0.9266, which is comparable to the other models.

5. Visualization

5.1. Residual Plots

Residual plots are used to assess the performance of the models. I create residual plots for the CatBoost model to visualize the differences between the predicted and actual values. The residual plot for CatBoost shows that the residuals are randomly distributed around zero, indicating a good fit. I also create an actual vs. predicted plot to visualize the relationship between the actual and predicted values.

5.2. Feature Importance

Feature importance plots help identify the most influential features in the model. I create a feature importance plot for the CatBoost model to visualize the importance of each feature. This plot helps in understanding which features contribute the most to the prediction of flood probabilities.

5.3. Q-Q Plots

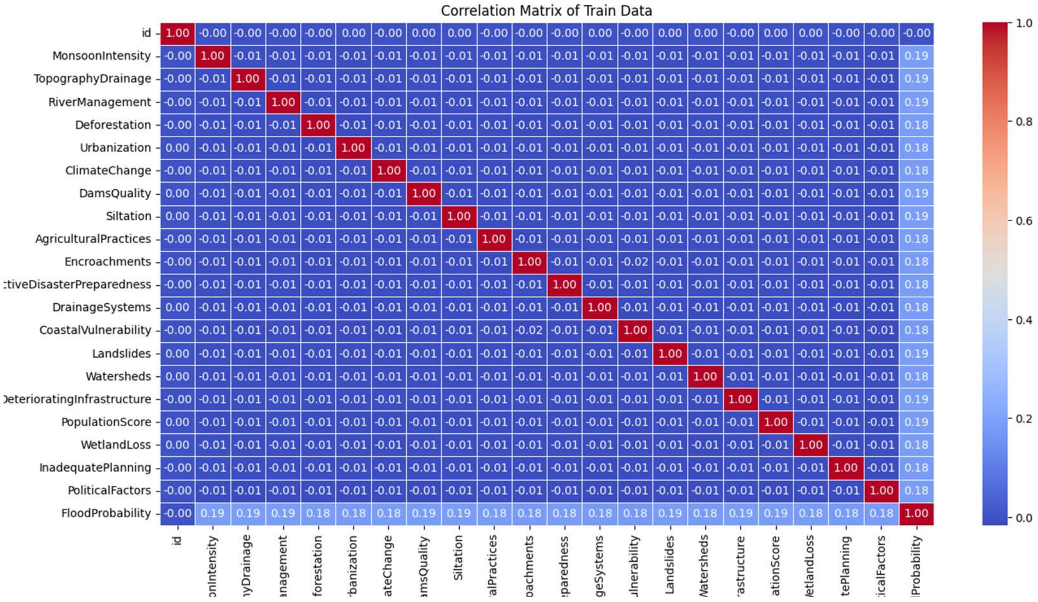
Q-Q plots are used to assess the normality of the residuals. I create Q-Q plots for the LightGBM and CatBoost models to visualize the distribution of the residuals. The Q-Q plots help in understanding whether the residuals follow a normal distribution, which is an important assumption for many statistical tests.

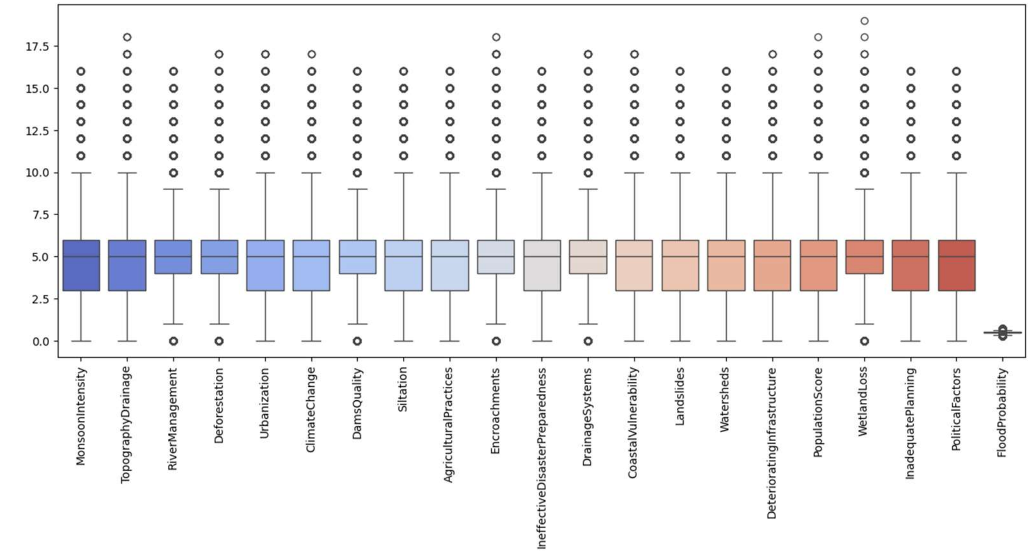
6. Predicted Values File

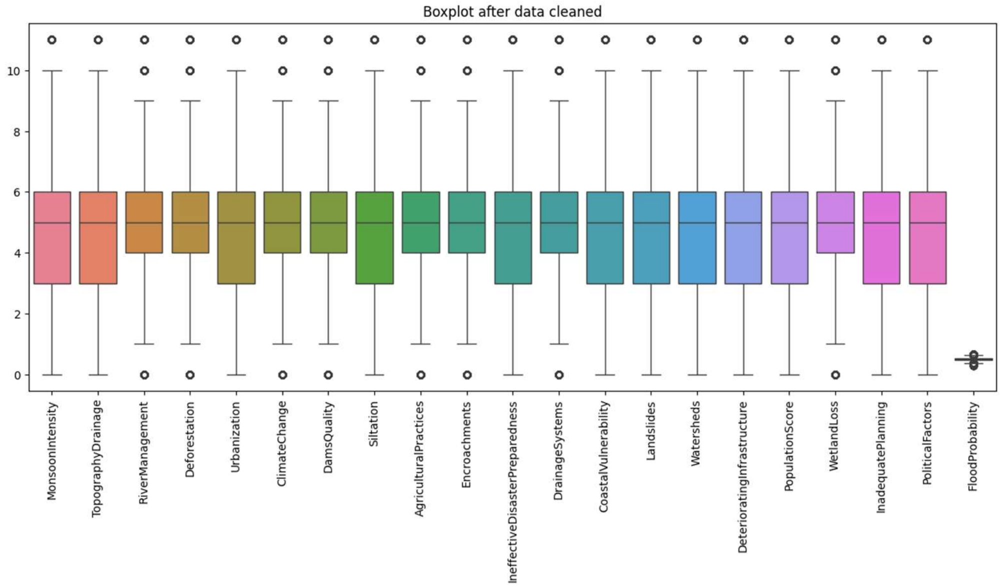
I create a DataFrame containing the predicted flood probabilities for the testing dataset. The predictions are a lighted average of the predictions from the LightGBM, CatBoost, and XGBoost models. I save the predicted values to a CSV file for further analysis. I also create a histogram to visualize the distribution of the predicted flood probabilities.

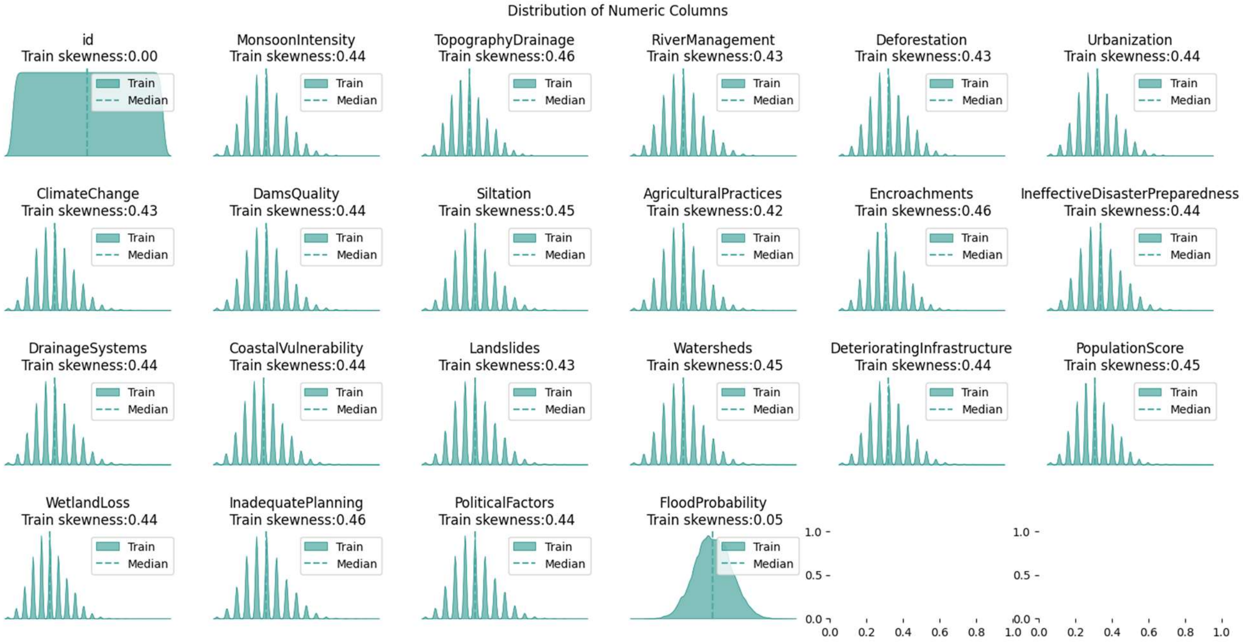
7. Conclusion

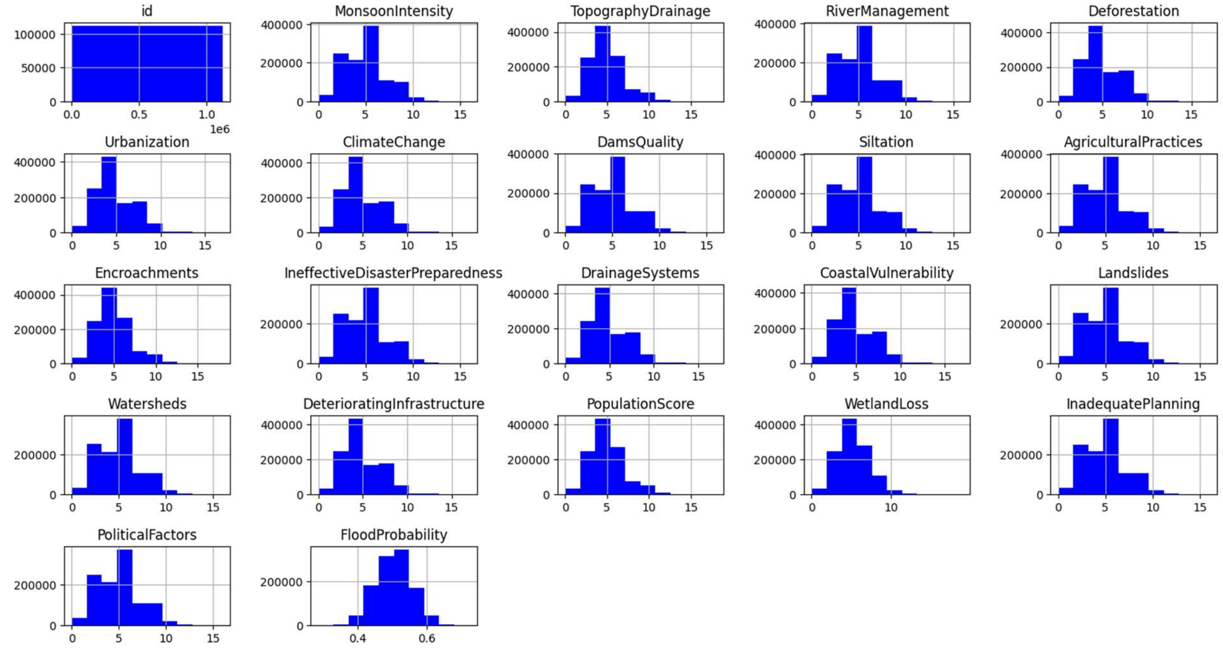
The analysis involved data preprocessing, feature engineering, model training, and evaluation. The CatBoost, LightGBM, and XGBoost models Ire trained and evaluated, with all models achieving high R-squared values. The residual plots, feature importance plots, and Q-Q plots provided insights into the model performance and the distribution of the predictions. The predicted flood probabilities Ire saved to a CSV file for further analysis. This project demonstrates the effectiveness of machine learning models in predicting flood probabilities, which can be crucial for disaster management and preparedness.

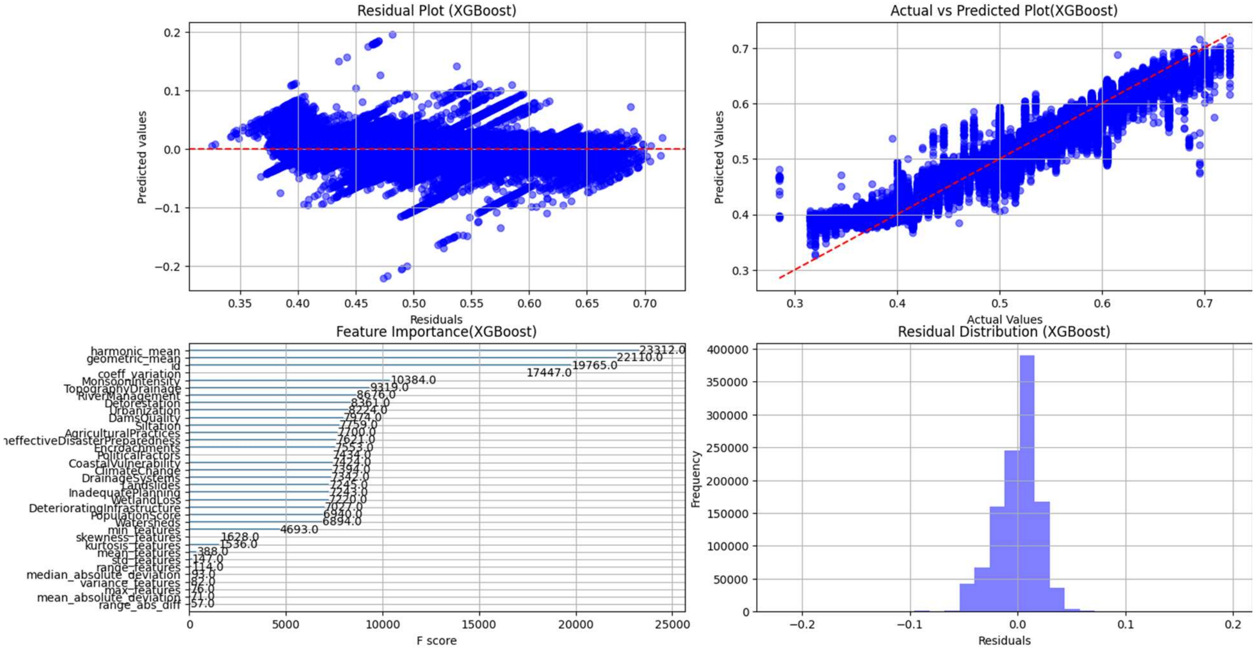

Correlation Matrix of Train Data
id 1.00 -0.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 -0.00
MonsoonIntensity -0.00 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
TopographyDrainage -0.00 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
RiverManagement -0.00 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
Deforestation -0.00 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
Urbanization 0.00 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
ClimateChange -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
DamsQuality 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
Siltation 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
AgriculturalPractices -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
Encroachments -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.02 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
:tiveDisasterPreparedness -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
DrainageSystems 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
CoastalVulnerability -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.02 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
Landslides 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.19
Watersheds 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 -0.01 0.18
DeterioratingInfrastructure -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 -0.01 0.19
PopulationScore 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 -0.01 0.19
WetlandLoss 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 -0.01 0.18
InadequatePlanning -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 -0.01 0.18
PoliticalFactors -0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 1.00 0.18
FloodProbability -0.00 0.19 0.19 0.19 0.18 0.18 0.18 0.19 0.19 0.18 0.18 0.18 0.18 0.18 0.19 0.18 0.19 0.19 0.18 0.18 0.18 1.00
id
onIntensity
nyDrainage
anagement
forestation
banization
ateChange
amsQuality
Siltation
alPractices
oachments
eparedness
geSystems
ulnerability
Landslides
Watersheds
astructure
ationScore
VetlandLoss
tePlanning
icalFactors
lProbability
1.0
0.8
0.6
0.4
0.2
0.0

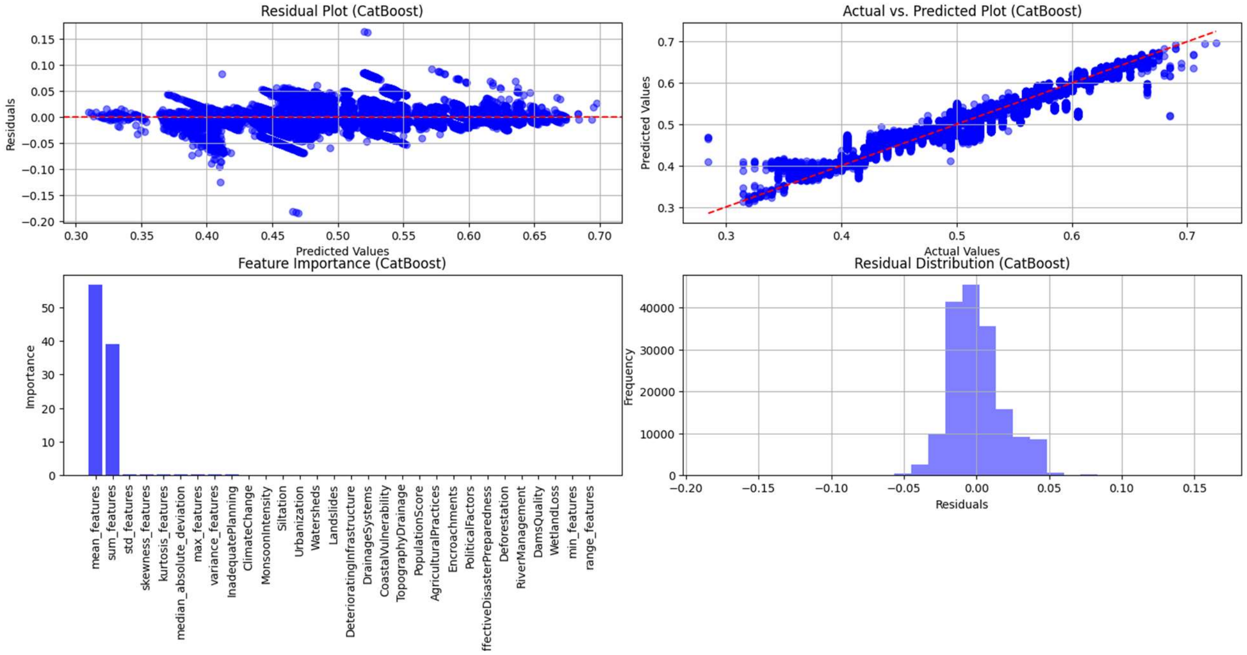

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
MonsoonIntensity
TopographyDrainage
RiverManagement
Deforestation
Urbanization
ClimateChange
DamsQuality
Siltation
AgriculturalPractices
Encroachments
IneffectiveDisasterPreparedness
DrainageSystems
CoastalVulnerability
Landslides
Watersheds
DeterioratingInfrastructure
PopulationScore
WetlandLoss
InadequatePlanning
PoliticalFactors
FloodProbability

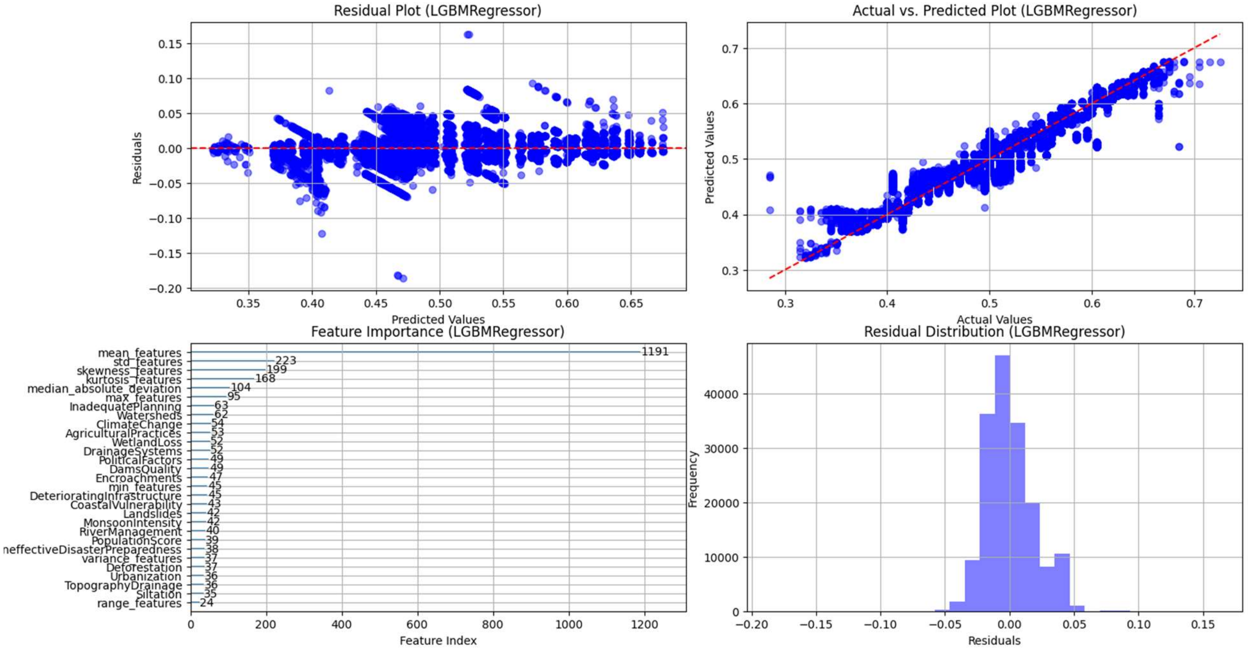

Boxplot after data cleaned
0
2
4
6
8
10
MonsoonIntensity
TopographyDrainage
RiverManagement
Deforestation
Urbanization
ClimateChange
DamsQuality
Siltation
AgriculturalPractices
Encroachments
IneffectiveDisasterPreparedness
DrainageSystems
CoastalVulnerability
Landslides
Watersheds
DeterioratingInfrastructure
PopulationScore
WetlandLoss
InadequatePlanning
PoliticalFactors
FloodProbability

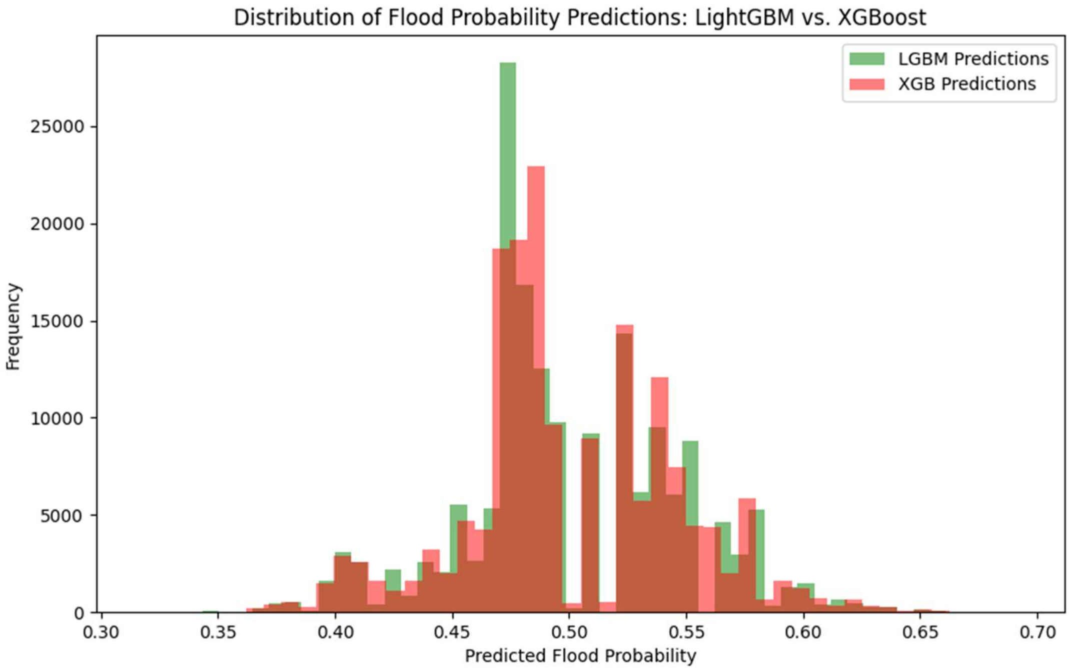

Distribution of Numeric Columns
id
Train skewness:0.00
MonsoonIntensity
Train skewness:0.44
TopographyDrainage
Train skewness:0.46
RiverManagement
Train skewness:0.43
Deforestation
Train skewness:0.43
Urbanization
Train skewness:0.44
ClimateChange
Train skewness:0.43
DamsQuality
Train skewness:0.44
Siltation
Train skewness:0.45
AgriculturalPractices
Train skewness:0.42
Encroachments
Train skewness:0.46
IneffectiveDisasterPreparedness
Train skewness:0.44
DrainageSystems
Train skewness:0.44
CoastalVulnerability
Train skewness:0.44
Landslides
Train skewness:0.43
Watersheds
Train skewness:0.45
DeterioratingInfrastructure
Train skewness:0.44
PopulationScore
Train skewness:0.45
WetlandLoss
Train skewness:0.44
InadequatePlanning
Train skewness:0.46
PoliticalFactors
Train skewness:0.44
FloodProbability
Train skewness:0.05
Train
Median
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

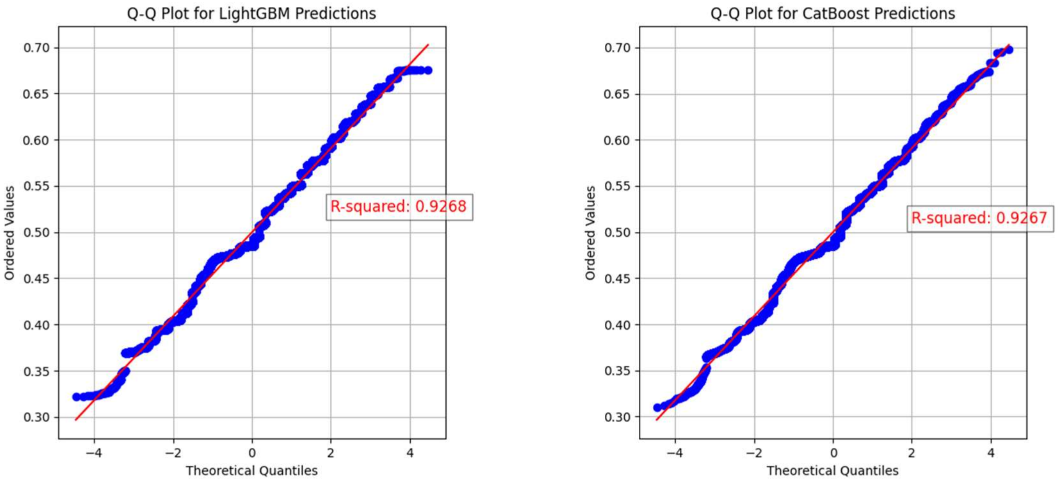

id
100000
50000
0
0.0
0.5
1.0
1e6
MonsoonIntensity
TopographyDrainage
RiverManagement
Deforestation
Urbanization
ClimateChange
DamsQuality
Siltation
AgriculturalPractices
Encroachments
IneffectiveDisasterPreparedness
DrainageSystems
CoastalVulnerability
Landslides
Watersheds
DeterioratingInfrastructure
PopulationScore
WetlandLoss
InadequatePlanning
PoliticalFactors
FloodProbability
400000
200000
0
5
10
15
0.4
0.6

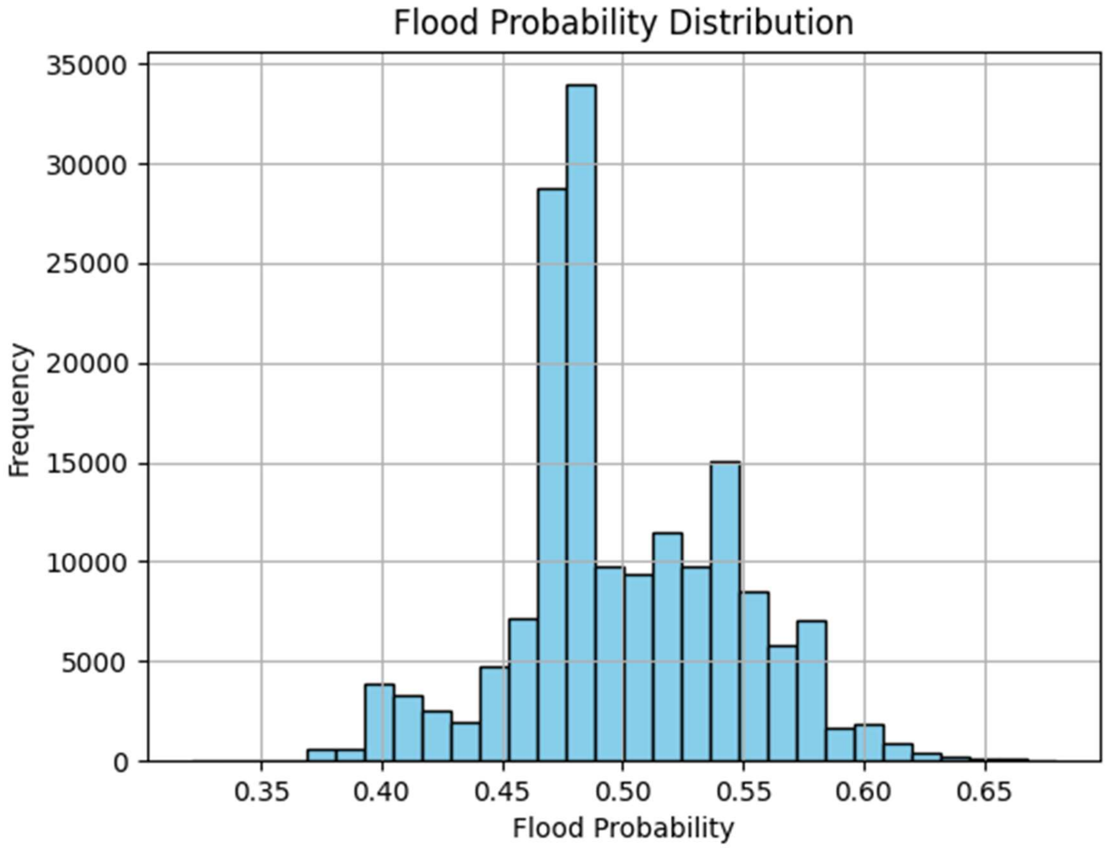

Residual Plot (XGBoost)
Predicted values
Residuals
0.2
0.1
0.0
−0.1
−0.2
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Actual vs Predicted Plot(XGBoost)
Predicted Values
Actual Values
0.7
0.6
0.5
0.4
0.3
Feature Importance(XGBoost)
harmonic_mean 23312.0
geometric_mean 22110.0
id 19765.0
coeff_variation 17447.0
MonsoonIntensity 10384.0
TopographyDrainage 9319.0
RiverManagement 8676.0
Deforestation 8361.0
Urbanization 8224.0
DamsQuality 7974.0
Siltation 7759.0
AgriculturalPractices 7700.0
neffectiveDisasterPreparedness 7621.0
Encroachments 7553.0
PoliticalFactors 7434.0
CoastalVulnerability 7424.0
ClimateChange 7394.0
DrainageSystems 7342.0
Landslides 7245.0
InadequatePlanning 7243.0
WetlandLoss 7220.0
DeterioratingInfrastructure 7027.0
PopulationScore 6940.0
Watersheds 6894.0
min_features 4693.0
skewness_features 1628.0
kurtosis_features 1536.0
mean_features 388.0
std_features 147.0
range_features 114.0
median_absolute_deviation 93.0
variance_features 82.0
max_features 76.0
mean_absolute_deviation 71.0
range_abs_diff 57.0
0
5000
10000
15000
20000
25000
F score
Residual Distribution (XGBoost)
Frequency
Residuals
400000
350000
300000
250000
200000
150000
100000
50000
0
−0.2
−0.1
0.0
0.1
0.2


Residual Plot (CatBoost)
Residuals
Predicted Values
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Actual vs. Predicted Plot (CatBoost)
Predicted Values
Actual Values
0.7
0.6
0.5
0.4
0.3
Feature Importance (CatBoost)
Importance
50
40
30
20
10
0
mean_features
sum_features
std_features
skewness_features
kurtosis_features
median_absolute_deviation
max_features
variance_features
InadequatePlanning
ClimateChange
MonsoonIntensity
Siltation
Urbanization
Watersheds
Landslides
DeterioratingInfrastructure
DrainageSystems
CoastalVulnerability
TopographyDrainage
PopulationScore
AgriculturalPractices
Encroachments
PoliticalFactors
ffectiveDisasterPreparedness
Deforestation
RiverManagement
DamsQuality
WetlandLoss
min_features
range_features
Residual Distribution (CatBoost)
Frequency
Residuals
40000
30000
20000
10000
0
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15


Residual Plot (LGBMRegressor)
Residuals
Predicted Values
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Actual vs. Predicted Plot (LGBMRegressor)
Predicted Values
Actual Values
0.7
0.6
0.5
0.4
0.3
0.3
0.4
0.5
0.6
0.7
Feature Importance (LGBMRegressor)
mean_features 1191
std_features 223
skewness_features 199
kurtosis_features 168
median_absolute_deviation 104
max_features 95
InadequatePlanning 63
Watersheds 62
ClimateChange 54
AgriculturalPractices 53
WetlandLoss 52
DrainageSystems 52
PoliticalFactors 49
DamsQuality 49
Encroachments 47
min_features 45
DeterioratingInfrastructure 45
CoastalVulnerability 43
Landslides 42
MonsoonIntensity 42
RiverManagement 40
PopulationScore 39
neffectiveDisasterPreparedness 38
variance_features 37
Deforestation 37
Urbanization 36
TopographyDrainage 36
Siltation 35
range_features 24
0
200
400
600
800
1000
1200
Feature Index
Residual Distribution (LGBMRegressor)
Frequency
Residuals
40000
30000
20000
10000
0
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15


Distribution of Flood Probability Predictions: LightGBM vs. XGBoost
LGBM Predictions
XGB Predictions
Frequency
25000
20000
15000
10000
5000
0
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Predicted Flood Probability


Q-Q Plot for LightGBM Predictions
Ordered Values
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
R-squared: 0.9268
−4
−2
0
2
4
Theoretical Quantiles
Q-Q Plot for CatBoost Predictions
Ordered Values
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
R-squared: 0.9267
−4
−2
0
2
4
Theoretical Quantiles


Flood Probability Distribution
Frequency
35000
30000
25000
20000
15000
10000
5000
0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Flood Probability