Flood Predictive Analysis Report

case study: Lagos State

Introduction

Recent incidents of flooding in Lagos State have raised significant concerns regarding the substantial losses and damages that could have been mitigated or prevented with early warning systems. Utilizing data analysis, we can predict potential flood events, enabling better preparedness and response measures. In light of these recent events, I have been tasked with leveraging my analytical skills and tools to predict the date of the next flood in Lagos.

Data Overview

i) Precipitation Dataset Overview

This weather dataset is crucial for predicting heavy rainfall and potential flooding. The key columns and their importance are:

- **datetime**: Aligns weather data with flood incidents.
- **precip**: Measures rainfall amount; >7.6 mm/hour indicates heavy rain..
- **precipprob**: Probability of rain; >70% suggests significant rainfall.
- **humidity**: Percentage of humidity; >80% indicates high moisture levels.
- **severerisk**: Risk of severe weather; high scores indicate potential for heavy rain

Other data in this dataset are: information on temperature, wind, solar and moon phases.

ii) Niger Delta Relative Lake Height Data Overview

The relative lake height dataset tracks lake height variations, which are crucial for detecting and forecasting floods. Key columns include:

- **date**: To track lake height variations over time. Which is crucial for identifying flood patterns and trends.
- **height variations**: Measurements of height changes. Essential for detecting and forecasting floods.

Other data in this dataset are: hour of the day, minute of the day, information on satellite station and instrumentation.

iii) NEMA Flood Data in Lagos State 2022 Data Overview

This dataset provides historical data on flood incidents in Lagos in 2022. Key columns:

- **state, LGA, community**: geographical identifiers.
- **date of occurrence**: Date of flood events.

Methodology

Data Collection and Preparation

1. **Data Sources**:
    - Precipitation data was sourced from weather datasets(VisualCrossing)
    - Lake height data was obtained from Nasa satellite measurements (Niger inner delta wetland)
    - Historical flood data was collected from NEMA reports
2. **Data Cleaning**:
    - Cleaning: I conducted data cleaning processes to address issues such as missing values, duplicate records, and inconsistent formatting in datetime fields before loading my data. This ensured the integrity and reliability of the datasets used for analysis.
3. **Feature Engineering:**
    - performed feature engineering to create new features that would enhance the predictive power of the model. This included generating indicators for heavy rainfall, and transforming qualitative indicators (e.g., incidence of flood into numerical representations.

```python
import pandas as pd

# Load datasets
weather_data = pd.read_csv('HNG task 2/Precip Data - 2022.csv')
lake_height_data = pd.read_csv('HNG task 2/Relative lake height niger-delta - wetland_time_series.csv')
flood_data = pd.read_csv('HNG task 2/2022 NEMA flood data - 2022 NEMA Flood Data.csv')

# Clean data
weather_data.dropna(inplace=True)
lake_height_data.dropna(inplace=True)
flood_data.dropna(inplace=True)

# Convert date columns to datetime with specified format
weather_data['datetime'] = pd.to_datetime(weather_data['datetime'], dayfirst=True, errors='coerce')
lake_height_data['datetime'] = pd.to_datetime(lake_height_data['datetime'], dayfirst=True, errors='coerce')
flood_data['DATE OF OCCURRENCE'] = pd.to_datetime(flood_data['DATE OF OCCURRENCE'], dayfirst=True, errors='coerce')

# Verify conversion
print(weather_data['datetime'].head())
print(lake_height_data['datetime'].head())
print(flood_data['DATE OF OCCURRENCE'].head())


# Create binary indicator for heavy rain days
weather_data['heavy_rain'] = weather_data['precip'] > 7.6

# Calculate moving averages of lake height
lake_height_data['lake_height_ma'] = lake_height_data['Height variation (meters)'].rolling(window=7).mean()
```

Data Collection & Preparation Process

Data Integration

1. **Merge Datasets:**
    - To consolidate relevant information, I integrated the cleaned weather data, lake height data, and historical flood incident data into a unified dataset. This integration allowed for comprehensive analysis and correlation between different variables contributing to flood occurrences in Lagos.

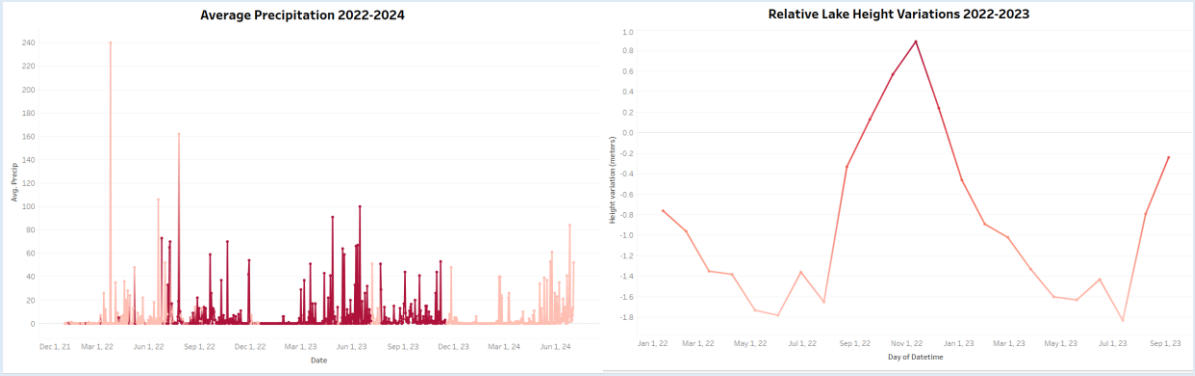
```python
# Merge weather data and lake height data on datetime
merged_data = pd.merge(weather_data, lake_height_data, on='datetime', how='inner')

# Align flood data by creating a binary indicator for flood days
flood_data['flood_occurred'] = 1
flood_data = flood_data[['DATE OF OCCURRENCE', 'flood_occurred']]
flood_data.rename(columns={'DATE OF OCCURRENCE': 'datetime'}, inplace=True)

# Merge with the main dataset
merged_data = pd.merge(merged_data, flood_data, on='datetime', how='left')
merged_data['flood_occurred'].fillna(0, inplace=True)
```

Merging the datasets

Exploratory Data Analysis (EDA)

- **Visual Exploration**: I conducted exploratory data analysis to gain insights into the distributions, relationships, and trends within the integrated dataset.

- **Statistical Analysis**: Utilizing statistical measures and visualizations, I identified key variables such as precipitation levels, humidity, and historical flood frequencies that significantly influence flood predictions.
- **Correlation Analysis**: I examined correlations between variables to understand their impact on flood occurrences, ensuring that only relevant and influential features were selected for model training.



Exploratory Data Analysis of the Datasets

Data Splitting

1. **Training Data**: For training the predictive model, I utilized data collected from January 2022 to December 2023. This time frame was carefully selected to cover a diverse range of weather conditions and flood incidents while ensuring the relevance of the information used for model development.
2. **Test Data**: I Reserved data from January 2024 onwards for testing and validating the model's predictive accuracy.

Model Building

- **Model Selection**: Based on my exploratory findings, I chose the Random Forest Classifier due to its ability to handle complex relationships and provide robust predictions for categorical outcomes like flood occurrences.
- **Feature Selection**: Using feature importance rankings from the Random Forest model, I identified the most significant predictors of floods and refined the feature set accordingly.

- **Training**: The model was trained using the prepared dataset from January 2022 to December 2023, ensuring it learned from a diverse range of weather conditions and historical flood incidents.

Model Validation and Testing

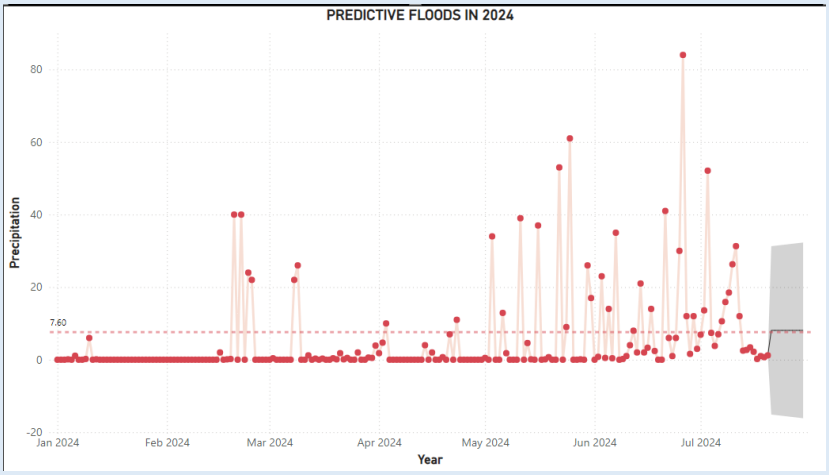
1. **Cross-Validation**:
    - Implemented k-fold cross-validation (k=5) to validate the model performance on different subsets of data.
    - Evaluated metrics such as accuracy, precision, recall, and F1-score.
2. **Testing**:
    - Applied the trained model to the test data from 2024.
    - Compared predicted flood dates against actual flood occurrences to assess predictive accuracy.

```
# Using yor  future weather data for 2024
future_weather_data = pd.read_csv('HNG task 2/2024 TEST MODEL - 2024 (1).csv')

# Preprocess future data similarly
future_weather_data['datetime'] = pd.to_datetime(future_weather_data['datetime'])
future_weather_data['heavy_rain'] = future_weather_data['precip'] > 7.6
future_X = future_weather_data[features_without_lake_height]

future_predictions = model_no_lake_height.predict(future_X)
future_weather_data['flood_prediction'] = future_predictions
```

Prediction

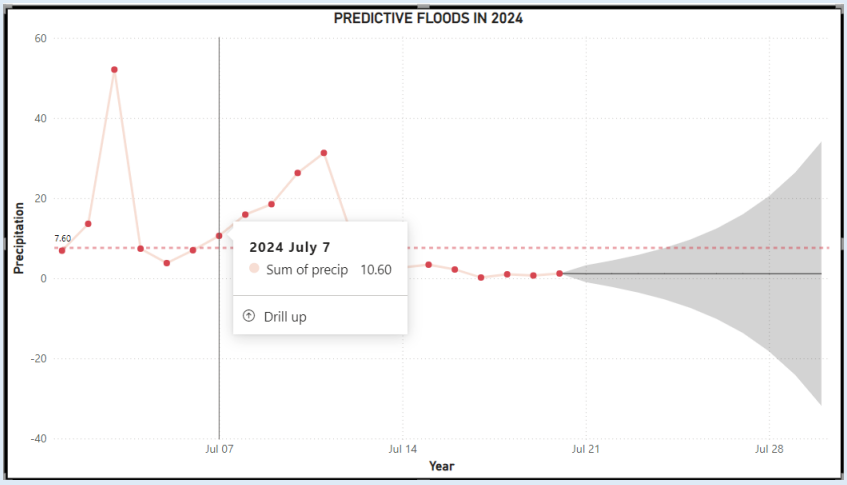
Prediction

1. **Future Predictions**: Using the trained Random Forest model, I made predictions for potential flood dates in 2024 based on weather forecasts and historical trends observed in the training dataset.



Flood Forecast of 2024

2. **My Model Predicted that the next flood will be on the 7th of July 2024 and for the next few days after.**



1A closer look at the prediction

Challenges/Setbacks

1. **Insufficient or Incorrect Data Source**: Some data sources were either insufficient or incorrect, impacting the analysis.
2. **Inadequate Structured Data on Flood Reports:** Sufficient structured data on flood reports in Lagos State is not readily available to the public.

Conclusion

Using the precipitation dataset and historical flood data, we can predict potential flood occurrences in Lagos. The model identifies heavy rainfall and associated weather conditions as key indicators. Despite the challenges of data quality and availability, the analysis provides a robust framework for flood prediction, helping in proactive flood management and preparedness.