Data Compilation Strategy for Wildfire Risk Prediction

Table of Contents

---

1. Datasets Overview

Historical Wildfire Data

- **MODIS (Moderate Resolution Imaging Spectroradiometer) Fire Data**: Provides historical fire data detected by the MODIS sensors on NASA's Terra and Aqua satellites.

Weather Data

- **NOAA's National Weather Service API**: Provides real-time weather data, historical weather observations, forecasts, and climate datasets.

Vegetation and Land Cover Data

- **USGS National Land Cover Database (NLCD)**: Offers detailed data on land cover categories, which can help in assessing the types of vegetation present and their susceptibility to fire.

Topographical Data

- **USGS Elevation Data**: This data includes elevation and terrain features, which are crucial for modeling how fires spread.

Climate Change Models

- **Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre**: Provides scenarios of future climate conditions which can be used to simulate future wildfire risks under different climate change scenarios.

2. Data Preprocessing Steps

- Cleaning
- Normalization
- Handling missing data
- Date formatting, etc.

3. Feature Engineering and Selection

Feature Creation from Compiled Datasets

We focus on extracting features that are relevant to predicting wildfire risks, such as:

- **Weather Derived Features**: We will create features like average temperature, humidity levels, and cumulative rainfall over the days leading up to the current date. This helps in understanding weather patterns that may contribute to fire risks.
- **Historical Fire Patterns**: We will analyze the frequency and intensity of past wildfires in different regions to create features that reflect the historical susceptibility of areas to wildfires.
- **Vegetation and Land Features**: From our vegetation data, we will calculate indices such as the Normalized Difference Vegetation Index (NDVI), which can indicate the health of vegetation and its susceptibility to catching fire.
- **Topographical Features**: Elevation, slope, and aspect derived from topographical data help in understanding how topography influences fire spread.

Feature Selection for the Model

Once we have engineered our features, selecting the most impactful ones is crucial to ensure model efficiency and effectiveness. We will employ several techniques and tools for feature selection:

- **Correlation Analysis**: We will first perform a correlation analysis to identify and remove highly correlated features, as they can introduce multicollinearity in our predictive models.
- **Importance Ranking from Ensemble Models**: Tools like Random Forest and XGBoost provide feature importance metrics which help in understanding which features contribute most to the model's predictions. We will use these metrics to prioritize features.

- **Recursive Feature Elimination (RFE)**: We will use RFE, a backward selection technique, to systematically remove features and evaluate model performance incrementally. This helps in identifying a subset of features that yields the best performance.
- **Principal Component Analysis (PCA)**: For datasets with high dimensionality, we will apply PCA to reduce the dimensionality while retaining the most informative features. This is particularly useful in preserving essential information while simplifying the model.

Tools Used for Feature Engineering and Selection

- **Python Libraries**: We will extensively use pandas for data manipulation, numpy for numerical operations, and scikit-learn for implementing machine learning methods like RFE and PCA.
- **Visualization Tools**: Tools like matplotlib and seaborn are employed to visualize the distribution of features and their relationships, aiding in our feature selection process.

4. Data Integration

Integrating Different Datasets

For our wildfire risk prediction model, integrating various datasets effectively is crucial to ensure comprehensive analysis. Here's how we plan to integrate our diverse datasets:

- **Common Key Identification**: We first identify common keys across datasets, such as geographic identifiers (e.g., ZIP codes, county names) and dates. These keys will serve as the basis for merging data from different sources.
- **Data Alignment**: We could align data temporally and spatially. For instance, weather data collected daily is aligned with historical wildfire occurrences recorded on specific dates, ensuring that the datasets correspond accurately in both time and space.
- **Consolidation**: Using data manipulation tools, we will consolidate datasets into a single structured format. This involves normalizing data scales, interpolating missing values, and resolving inconsistencies in data formats and units.
- **Feature Integration**: We will integrate features derived from different datasets, such as weather conditions, vegetation indices, and topographical features, into a unified dataset that feeds into our predictive models.

Tools Used for Integration

- **Data Integration Tools**: We will use Python libraries like Pandas for data manipulation, along with ETL (Extract, Transform, Load) tools like Apache Nifi or Talend for automating data integration workflows.