Wildfire Prediction Based on Environmental Factors
Milestone

Member 1	Guangkai Chen	guangkai22@vt.edu
Member 2	Kaiyi Chen	kennychen@vt.edu
Member 3	Chujia Chen	cjchen@vt.edu

Introduction

In recent years, the world has witnessed a surge in the frequency and intensity of wildfires. These catastrophic events, driven by a myriad of environmental factors, have wreaked havoc on ecosystems, displaced communities, and caused billions in damages. The unpredictability and rapid onset of these fires have made them especially challenging to manage. However, with the advancements in data science and machine learning, there lies an opportunity to harness vast datasets to predict these occurrences, facilitating early preventive measures. Armed with the promise of data science and machine learning, our team set forth on a mission to construct a predictive model for wildfire occurrences. This report provides a comprehensive overview of the project's progress, from data acquisition to preliminary modeling, as we edge closer to the final phase of our endeavor.

Achievements

Data Collection

We have successfully gathered the required datasets to fuel our predictive model. Our primary sources include reputable institutions such as NREL and NOAA:

- Solar Resource Maps from NREL
- Solar datasets from NREL
- Comparative Climatic Data from NOAA NCEI
- Daily Climate Data from NOAA NCDC
- Precipitation data from Weather.gov
- Climate Data from NOAA
- Further datasets from NOAA and USGS as necessary
- Data Processing & Inspection

After consolidating the data, our preliminary estimatation suggest that our dataset consists of several hundred thousand records, each representing specific geographical locations and timeframes. This rich and resourceful dataset will provide a comprehensive and valuable insights of environmental attributes juxtaposed with wildfire occurrences.

We also preformed a deep inspection of our data as this action revealed its sparse nature. This sparsity will be a critical aspect to consider during preprocessing and modeling, as it can affect the performance and reliability of our predictive algorithms.

Preprocessing Steps Completed

- *Data Cleaning*: We have addressed inconsistencies in the data, handled missing values, and eliminated outliers to ensure the data's integrity.
- *Data Merging*: Data from the aforementioned sources has been successfully merged to create a unified dataset, paving the way for further analysis.
- *Feature Engineering*: Preliminary steps have been taken to generate new features that might be relevant for prediction, ensuring a rich and comprehensive feature set.

Data Teasing and Feature Extraction

Having delved deep into the data, we have successfully teased out critical insights. Several features have been identified and extracted, positioning us advantageously for the experimental phase. Our feature set, estimated to span 50-100 variables, encompasses:

- Standard environmental indicators like temperature, humidity, and rainfall.
- More intricate factors such as solar radiation, wind patterns, historical fire incidents, and other environmental markers.

Preliminary Modeling

Guided by our proposal's blueprint, we initiated the modeling process on a small subset of our dataset. This preliminary phase allowed us to gauge the potential challenges and refine our approach. Initial experiments with Decision Trees and Logistic Regression have provided valuable feedback, prompting us to fine-tune our strategies for the main dataset.

Challenges Faced

- *Data Sparsity*: Our primary challenge remains the sparse nature of our dataset. Addressing this will be pivotal, as sparsity can adversely impact model accuracy and generalization.

Next Steps

- Data Normalization & Transformation: We will scale features to ensure they have equal significance in models and convert non-numeric data into a format suitable for machine learning algorithms.
- Modeling: With the preprocessing steps nearing completion, we will delve into building our predictive models. As outlined in our proposal, we'll be exploring Decision Trees, Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting Machines.
- Addressing Data Sparsity: Given the sparse nature of our data, we might explore techniques such as data augmentation or specific algorithms tailored for sparse data. Additionally, methods like Principal Component Analysis (PCA) may be employed for dimensionality reduction if required.
- Model Evaluation: Once our models are developed, rigorous evaluation will be conducted to assess their predictive power. This will ensure the reliability of our predictions.

Conclusion

The Wild-fire Prediction Project is progressing steadily. We have achieved significant milestones in data collection and preprocessing. As we transition into the modeling phase, we remain committed to delivering a solution that will be useful in predicting and mitigating the devastating effects of wildfires. We are optimistic about the future milestones and the eventual success of this project.