Team SAAW: Fire Predictor/Wildfire Project

Project Description:

Our project aims to build a model that can predict how likely a specific point of interest will have a wildfire outbreak. More specifically, we are aiming to build a model that can output a number between 0 and 1 to show the likelihood of a wildfire occurring.

Utilizing wildfire data (2017-2022) from a region in Canada, we were able to train an XGBoost Model to categorize 6 levels of fire risk, from “No Risk” to “Extreme Risk” based on previous fire occurrences and weather data fitting the classification.

In our front-end minimum viable product, the user is greeted with a map of active wildfires in Canada. The user can then input a location and the Streamlit outputs the associated current wildfire risk, through our XGBoost machine learning model.

Our ultimate goal is to test in multiple high-risk regions in North America (California, for example) to make our predictions more accurate. As of now, the analysis is based on the current daily weather features of the location, but we plan on incorporating a one-week forecast so that users can view their current and one-week forecast risk.

We want our map to be more interactive and highlight areas in which we have collected data and performed XGBoost classification to enhance accuracy for fire prediction within the region.

We also want to add a menu with features like an interactive time series of forest fire history, historic wildfire density maps, and more folium-based visualizations.

MVP
https://colab.research.google.com/drive/1IPe2Cb7w3TXg-IkbvOFf4-5YJZAz_Dlv?usp=sharing
(Two HTML files need to be uploaded every time the notebook is ran and since we were unable to find a way to input the files permanently, here is a video showing how our website currently works: video)

ML Models Performance
https://colab.research.google.com/drive/1Q8dYHSA8nq1nTvJmVjf1PRtX3ahe3lJi?authuser=2

Data processing Google Colab link:
https://colab.research.google.com/drive/1vC1rBLxlaRajT060TrVRD4qPFUPUNnXs?usp=sharing

Machine Learning part of the Canadian data:
https://colab.research.google.com/drive/1J-mzsh3bDbgFDpIlgeMCW-5UpUVYO4Sy?usp=sharing

Sample Testing Regions: Canada, United States, and Portugal

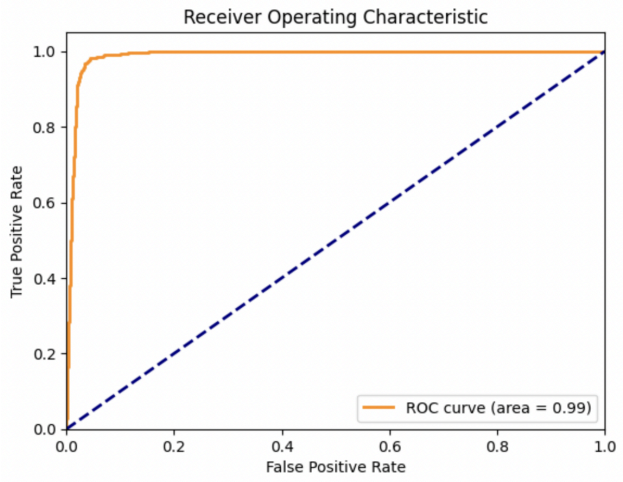
Weather API Variables:

Date
Weather_code
Temperature_2m_max
temperature _2m_min
Temperature_2m_mean
Apparent_temperature_max
Apparent_temperature_min
Apparent_temperature_mean
Daylight_duration
Sunshine_duration
Precipitation_sum
Rain_sum
Snowfall_sum
Precipitation_hours
Wind_speed_10m_max
Wind_gusts_10m_max

Wind_direction_10m_dominant
Et0_fao_evapotranspiration

Training the XGBoost ML Model:

- 1,000+ data samples (days with/without wildfire) with a random 100,000 sample of "no-fire" and all 7,000 instances of "fire"
- Utilized XGBoost classifier (using pickle dumps to change the model format to binary, and using predict_proba and then transforming the scores to be evaluated under categorization thresholds (through a z-score). If the score reaches the top percentiles of the dataset, the higher its risk factor is
    - ```
      percentiles = [85, 96, 99, 99.6, 99.9]
      ```



Additional Areas to Explore:

- Wind gust vectors for fire spread to different locations, adding to fire risk for places that would otherwise have lower weather risks
    - Using wildfire power to estimate magnitude and growth
- Fit Canadian dataset to transform to US dataset
- Predicting time-series for wildfire risk
- Any other predictive features, such as human population, past fire, tree density (perhaps not accessible for a student, but might be more accessible for professional forest managers)
      ```

Contributions:

Alex:

Past: Facilitating product management, United States Active Wildfire API, Trained a sample model using machine learning models for a dataset in a Portuguese National Park, Milestone 4 document, Assisted with multi-variate regression on the Canadian dataset, Assessed weather data features, Organized product EPICS

Milestone 5: Tested various ML models to find optimal classifier, worked with Sicheng and TA's on the model selection and data imbalance, Talked to Emiel Cox (CEO of satellite imaging carbon sequestration company, Genvision) about data imbalances and ML models, augmented data frame for machine learning testing, assisted with XGBoost model and spearheaded classification, connected model to front-end Streamlit, managed product sprint, wrote milestone V writeup, enhanced streamlit through formatting the predictor, including the bar plot

Future: Regional data and model expansion in North America, enhancing Streamlit user interface and experience

Ahhyun:

Past:
helped to collect data/apis, collected live fire data that could be used for machine learning, built majority of Streamlit MVP
Future:
work on the front-end aspects of the project, presentation

Sicheng:

Past:

Helped to collect data on Canadian as well as other countries in the world's historical fire data;
Helped to improve the code to access weather data from various location and various dates based on William's data
Helped to write tutorials and functions for saving, loading, deleting, changing data in SQL database

Analyze, visualize and process the Canadian fire data: visualize the fire density of different places in Canada and find the area with the highest fire density and select the region as the subject of the study; properly process the fire data and get a DataFrame that has the information of how many times fire occur decide the mass of the region and to how many digits after the decimal point of the coordinates we use to simultaneously ensure that database has enough data points (finally 7 million columns with 20 thousand columns with positive value of fire) and the space the DataFrame takes is not too large that the google colab collapse;
Helped to clean and concat data of weather: first tackle weather URL calling times limit by creating several copies of same google-colab notebooks and save the data to the database and call the url at different times of a day and finally put the data together into the same dataframe;
Helped to process the fire and weather data so that they have the same form and can be merged together ;
Helped to adapt the columns of the fireandweather DataFrame to the form that can be read by machine learning models and helped to process features so that the model can have more predictive power
Try some simple models like logistic regression but faced problem of not enough RAM space of google colab notebook
Facilitated working through the machine learning models and inputting balanced data into the XGBoost model

Future:
Work on making more machine learning models for each region

William:
Past - Helped get the weather API to access historical weather data, helped develop code that automates the process of getting weather data from the NASA API, helped clean up the SQL database

Milestone 4-5: Created the map and time series visualizations; created the front-end (Streamlit) prototype for milestone 5

Future - Help with the front-end development; help develop more visualizations that the user can use to explore the data that we gathered.