PLEASE SCROLL DOWN TO VIEW DELIVERABLE 2

DS/CMPSC 410 Mini-Project Deliverable 1: Activity Report

Project Title: Wildfires Big Data Analysis

- Team Members:
    - Alvaro Tapia
    - Sanidhya Singh
    - Chinmay Ghaskadbi
    - Jorge R Meza Cabrera
    - Ahmed Mosaad
    - Nakshatra Sharma
    - Evan Dougherty

Activities performed:

1. The first step for our coding process, aside from loading the data without the kernel crashing, was to preprocess the dataset (5.3GB) and perform feature importance and feature selection. For this purpose, we shrunk the dataset into one that only takes the first 100,000 rows (dataset of around 230MB) for efficiency purposes in order to avoid kernel crashing. We plan to later use the best features selected on the original dataset, the main reason we wanted to first tackle feature selection was due to the large number of columns/features we had available because trying to utilize each one within an ML model would not be entirely feasible. As such we will decide which certain columns/features to prioritize that would have a greater potential impact on the model.

Note: We do possess a 22 GB version of the dataset. However, for the convenience of uploading and running our initial trials, we are going to be running our code on the smaller versions of the data and once we are confident with our code, we shall run it with the larger dataset.

2. We also initiated our first attempt for applying a machine learning algorithm to the dataset which was lasso regression, as there are many variables we aren't positive on whether or not it is useful to the data/model. The HTML we submitted regarding lasso is currently bare-boned as we primarily wanted to use it as a method to help us utilize the naive method of feature selection.

3. After preprocessing the data and selecting the relevant features, we consider performing MapReduce when handling the large-scale dataset (22GB), so far we have developed code for the MapReduce algorithm but have not yet run it along with all the other code since right now we are using a small version of the dataset and

this can be an overkill cause it could introduce unnecessary overhead. This process is very important for the future since MapReduce will help us to handle this large dataset by performing parallel processing and increasing the speed of the analysis by making it scalable.

4. With the help of different libraries and sources available we decided to find potential algorithms to perform data visualization. For this first part, we created a correlation matrix algorithm so we can ensure that the features selected are the correct ones and that will potentially bring the most value to our investigation. In the same way, we developed some code that hasn't been run yet because we are still deciding if it will add relevant value to our investigation the code we did was for the creation of heated bar graphs in order to visualize and identify the impact specific features may have if they are studied separately, and also developed some linear models (lasso regression model) after filtering all features and choosing the most important ones for visualization that will be included in the final document report.

5. We also began to research and understand how to use correctly the persist() and cache() methods that we plan to integrate later in the investigation to increase the efficiency of our code and reduce the amount of run-time.

6. We are also actively looking at various techniques learned in recent previous labs and using any techniques that could be helpful for this project. Such as creating k-means instances and performing visualizations. Applying k-means clustering to group similar data points together. This can help identify patterns or regions where wildfires are more likely to occur based on the features provided. The clusters may represent areas with similar environmental conditions that could be conducive to wildfires.

7. Started implementing PCA to find the principal components (Investigate more about this).

DS/CMPSC 410 Mini-Project Deliverable 2: Activity Report

Project Title: Wildfires Big Data Analysis

- Team Members:
  - Alvaro Tapia
  - Sanidhya Singh
  - Chinmay Ghaskadbi
  - Jorge R Meza Cabrera
  - Ahmed Mosaad
  - Nakshatra Sharma
  - Evan Dougherty

Activities performed:

1. Principal Component Analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. As a first attempt with the creation of the PCA analysis code, we were able to make it run on the 100 MB dataset. After that, the same process was performed in the 5 GB dataset just to make sure it also worked on the official dataset and it effectively ran as expected.

2. Researched some possible ways that we could improve the efficiency of our PCA analysis. Also decided to begin researching and understanding how to correctly use the persist() and cache() methods that we plan to integrate soon in the investigation to potentially increase the efficiency of our code and reduce the amount of run-time.

3. As we were also transitioning from the smaller 200mb dataset to the 5gb version of the dataset we ran into various errors regarding various NaN/null values in a large number of columns for a specific count of 29 rows (in the 5gb), and since we have a large amount of observations we currently believe just removing those rows won't create a signifigant enough of an impact to worry about as we have a very large number of observations to begin with. This decision may change depending on the number of null values that appear in the observations within the large 20+gb dataset. While we could just use dropna() the other code we have can be used for future operations on the dataset as well.

4. After a discussion about what model to implement, we decided to discard using K-means and began to look more into Random Forest Regression modeling on our split dataset.

5. Random Forest regression is a flexible, easy-to-use machine learning algorithm that produces great results. A type of learning model where a group

of weak models combine to form a powerful model. In our case PCA analysis is successful, so Random Forest Regression can be used to predict a continuous outcome based on the principal components.

6. There is progress with using map reduce for using the bigger dataset. So far this is a rough draft of the code and will be improved in the future days as we progress with the other parts of the project.

7. Future Steps planned, since we were successful in implementing the PCA analysis we are short-listing some more ways to visualize the results. After visualizing and interpreting the results, we can draw conclusions about our data and start writing the final report document. This will involve identifying key variables that contribute to wildfires and predicting the severity of future wildfires based on certain conditions.

Link to GitHub: https://github.com/abt5572/DS410-Alpha

Link to Dataset: https://drive.google.com/file/d/1B582y8_cPWxNuevpm3ZM-SZf_23HRUAQ/view?usp=share_link

Link to Dataset's GitHub:
https://wildfire-modeling.github.io/