DS/CPSC 410 Project Deliverable 1: Project Plan and Preliminary Setup

Project Title: Wildfires Big Data Analysis

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

Team Composition

The whole team's plan is to collaborate effectively by constantly communicating our findings, results, and improvements in the project. We will also constantly meet because we all have to know what process to follow for the investigation. Meetings will be done at least once a week, in person preferably.

- Alvaro Tapia: Majoring in Applied Data Science, experience in Python machine learning techniques on developing models and usage of other methods. Also, experience in writing documents and projects. My responsibility in the investigation is to help develop machine learning models and methods such as feature importance, correlation matrix, feature selection, and train-test-validate models in order to find the best model (using PySpark). Also, I will explain the findings in the official report document. Will also support the development of PySpark algorithms such as hyperparameters, MapReduce, data management, and more.

- Chinmay Ghaskadbi: Majoring in Applied Data, my background is predominantly in using machine learning with anonymized medical data. My professional experience is working as a Data Scientist on a project with the American Red Cross as well as an engineering intern with GE Aerospace. My role in this project would be feature engineering as well as implement the machine learning model. Furthermore, I will be heavily involved in writing and development of the project paper.

- Sanidhya Singh: Majoring in Applied Data Science and as such I've taken classes working with multiple machine learning models and various datasets. While the majority of the datasets were quite small there were instances in my internships where I was able to learn about classification using larger datasets using models like Xgboost through services like AWS. My primary role will mainly involve hyperparameter testing during the development of this project.

- Jorge R Meza Cabrera: Majoring in Computer Science, I don't really have experience with data science or machine learning since this is the first data science class I have taken.

- Nakshatra Sharma: In my senior year doing Computer Science. This is my first data science class as I have more experience with computer science and computer

engineering classes. My professional experience has ranged from being a part of a management consulting firm McKinsey's data science team called Quantum Black. Further, I also have experience working with Penn State's legal team.

- Ahmed Mosaad: Majoring in Computational Data Science. I am a senior, and I have had a mix of classes like computational stats which is the math behind machine learning and other math and stat classes. This is my second machine learning class. I am currently taking another one that introduces the algorithms and basics of machine learning. I have experience with R and making prediction models and such, I also have had experience with using AI or chatbots to boost effectiveness in company workflow.
- Evan Dougherty: I am a senior majoring in computer science. I have taken CMPSC 448, which was about machine learning so I have some experience with that. I also worked with databases for my 2 internships (Veradigm and Peraton), but it was not extensive. Given that I have more experience on the ML side of things, I plan on helping to build the ML model.

Project Objectives

Within this project our goal is to predict the FPR score using ALS to determine the likelihood of a forest fire happening while also utilizing and improving our skills for data cleaning and preparation we acquired in class on a significantly larger dataset.

Dataset Selection

For the purpose of this investigation, our group selected the WildFireDB as the data source. This dataset is significant as it utilizes previous historical wildfire occurrences with relevant features extracted from satellite imagery to form an open-source wildfire dataset. This is the first open-source wildfire dataset that contains information about past wildfire occurrences and also contains data about features extracted from satellite imagery. It contains 17 million data points collected from 2012 and 2018. There is a small version of this dataset with a size of 5 GB and also a full version with a size of 30 GB. This dataset contains cumulative information from different websites and sources such as Landfire for topography, Visible Infrared Imaging Radiometer Suite (VIIRS) for thermal anomalies, Metostat for weather data, and some other raw data from the National Oceanic and Atmospheric Administration (NOAA).

Project Timeline

- 10/6/2023: First meeting in person to talk about our initial project proposal and start working on the document. Also organizing future meetings and roles that each of us is going to have.
- 10/12/2023: Planned to be the second meeting to develop an approach to tackling the problem and finalizing the dataset as well as finishing the project proposal document.
- 10/18/2023: Meet again either online or in person in order to start implementing feature engineering data algorithms. Also, fix or add any final details to the project proposal.

- 10/27/2023: Start development for the purpose of the investigation, get graphs, tables, images, and more; in order to later start the final report of the project.
- 11/10/2023: Finish code and implement optimization techniques
- 11/17/2023: Implement the findings and results from the code and begin writing the report.
- 11/24/2023: Finish and submit, before this, we should contact the professor to see if everything looks good.
- 11/27/2023 Submission

No issues found yet, it seems that our goal is very clear and the dataset is not hard to manage. We are still testing the small dataset (5GB) and in the future, we will start trying to implement the 30GB dataset.

Tools and Technologies

Besides the machine learning methods and model implementation that are going to be used for feature importance, classification, and selection, as well as finding accurate models for prediction, we are going to use the following features and algorithms from the PySpark library:

- Try to find the methods for feature importance selection and ml models to be used in Pyspark.

- Alternating Least Squares (ALS)
    - This is from the ml.recommendation package and we will use it to run a machine-learning model on our data in order to predict future wildfires.

- CheckPoint to improve the efficiency of iterative processing using Spark
- Machine Learning with Alternating Least Squares (ALS): Understanding of ALS for forest fire prediction.

- MLlib to implement machine learning models.

- Hyper-parameters identification for better evaluation of the chosen model with testing data.

- Utilize various Python libraries to establish a form of error evaluation during the evaluation of wildfires.

- Use persists and/or checkpoints to improve performance.

- If useful use what we learn in the future (Lab8, Lab9, Lab10).

Project Plan Summary

The overarching goal of our project is to utilize the WildfiresDB data and clean it so that it is then usable for potential feature analysis and wildfire prediction using machine learning. WildfiresDB is an open-source dataset that links wildfire occurrence with relevant features. These features extracted from satellite imagery include pieces relating to weather, vegetation, and the canopy which

The project aims to predict the Forest Fire Probability Rating using Alternating Least Squares. This prediction will help determine the likelihood of a forest fire occurrence. Additionally, the project seeks to enhance skills in data cleaning and preparation on a larger dataset compared to what was covered in class.

In order to complete this project, we will have to improve our skills in some areas:

- Improved data-cleaning knowledge
- Greater Understanding of how to use ALS
- Collaboration
- Project Management
- Communication and Documentation
- Understand the computation behind our models
    - Numerical literacy to understand the dataset

By working on these skills, the team will be well-equipped to clean the WildfiresDB data, perform feature analysis, and build an accurate wildfire prediction model using ALS. This will not only lead to a successful project outcome but also enhance the team's capabilities in data science and machine learning.