Formal Statement Problem: Predicting Burn Areas of Forest Fire in Northern Portugal

Alison Jing Huang

Objective and Motivation

Forest fires are devastating disasters that can burn acres of land and engulf everything within minutes. According to *National Geographic*, each year on average more than 100,000 forest fires clear 4 million to 5 million acres (1.6 million to 2 million hectares) of land in the U.S. Wildfires have burned up to 9 million acres of land and a single one can move at speed of up to 14 miles an hour (23 kilometers an hour) destroying vegetations, homes, livestocks and human lives.The most recent October 2017 Sonoma County fire namely **TUBBS** (currently still under investigation) affected 36,807 acres, destroyed 5,643 structures and killed 22 people, is the single most destructive wildfire in California State history.According to L.A Times, TUBBS fires caused at least $3 billion in insured losses and this number can go up as high as $65 billions. Therefore, **fast fire detection and prediction is key element for controlling and preventing such critical environemntal issue, specifically the ability of accurately predicting the burned area of forest fires can significally help optimize fire management efforts and reduce damages and losses.**

The purpose of this project is to examine a pre-established forest fires dataset of Montesinho Natural Park of Northern Portugal dated between January 2000 to December 2003. This study will help **CAL FIRE** (California Department of Forestry and Fire Protection) gauge and assess the different attributes/variables that may contribute to the causes and sizes of fires.

Background

According to The Associated Pressed(**AP**) reported last year, half of California's $469 million emergency fund for large wildfires has already been solely used in the three months period of firefighting costs.Even before the sudden explosion of wildfires in California, federal wildfire suppression costs had already skyrocketed to unprecendented levels to as high as $2.4 billon in 2017, pulling funds out of fire prevention related projects.The Department of Agriculture(USDA) projected that the projected continued growth in the 10 year-average cost of fire suppression through 2025 is rising to nearly $1.8 billion, which amounts to nearly $700 million decrease in non-fire program funding in the next 10 years.

In order to reduce fire suppression costs, data mining and using automated machine learning tools to analyze the raw data and extract high-level information can bring valulable insight for the decision makers to take appropriate measurements and develop strategies to reduce firefighting time and cost.

As part of **CAL FIRE** team, the Office of the State Marshal(OSFM) reached out to the fire prevention engineering program at UC Berkeley in collaboration of **Cal-Adapt Project group** to tackle such critical issue, and hired me as a Data Scientist Intern to develop ML models to predict forest fires using Meterological data provided by UC Irvine Machine Learning Depository webpage as a preliminary study of fire prevention.

This study will consider fire data from the Montesinho natural park, from the Tr´as-os-Montes northeast region of Portugal. This park contains a high flora and fauna diversity. Inserted within a supra-Mediterranean climate, the average annual temperature is within the range of 8 to 12 Celsius.The data used in the experiments was collected from January 2000 to December 2003 and it was built using two sources. The first database was collected by the inspector that was responsible for the Montesinho fire occurrences. At a daily basis, every time a forest fire occurred, several features were registered, such as the time, date, spatial location within a 9×9 grid, the type of vegetation involved, the six components of the FWI system and the total burned area. The second database was collected by the Braganc¸a Polytechnic Institute, containing several weather observations (e.g. wind speed) that were recorded with a 30 minute period by a meteorological station located

in the center of the Montesinho park. The two databases were stored in tens of individual spreadsheets,under distinct formats, and a substantial manual effort was performed to integrate them into a single dataset with a total of 517 entries.

Reference:

http://archive.ics.uci.edu/ml/datasets/Forest+Fires

http://www3.dsi.uminho.pt/pcortez/fires.pdf

http://calfire.ca.gov/index

http://www.fire.ca.gov/communications/downloads/fact_sheets/Top20_Destruction.pdf

https://www.iii.org/fact-statistic/facts-statistics-wildfires

https://www.fs.fed.us/sites/default/files/2015-Rising-Cost-Wildfire-Operations.pdf

Stackholders (hypothetical)

- **Alison Jing Huang** - Data Scientist and Project Modeler for Berkeley Fire Prevention Engineering team & CAL FIRE R&D Program. She is responsible for throughly studying the dataset and develop predicitve modeling to identify the important meterological attributes associated with forest fires area.
- **J. Keith Gilless** - Chair of Board of Forestry and Fire Protection, Dean & Professor of UC Berkeley *Department of Environmental Science, Policy and Management.* He is the lead subject matter expert and Senior Advisor to Governor Jerry Brown. Based on the predicted model, he will give appropriate recommendations to the state legislature on fire prevention and wildland protections.
- **Ken Pimlott** - CAL Fire Director, under his direction, CAL FIRE will implement appropriate statewide Fire and Emergency Responses and improve education and safety trainings around the strongely identified attributes linked to forest fires.
- **Cal-Adapt** - is a site developed by UC Berkeley's Geospatial Innovation Facility(GIF) with funding and advisory oversight by the California Energy Commission's Public Interest Energy Research (PIER) Program. They will use the project result to further collaborate with CAL FIRE and implement visualization tools to study fire areas predictions.
- **U.S Department of Agriculture** - provides financial statements, budget and costs associated with fire suppression, and collaborate with CAL-FIRE to closely monitor the costs and compare the statistics throughout the next few years to see whether the implemented plan based on this project can lower the costs.

Success Metrics

- One of the success metrics is to have strong meterological attributes linking to forest fires burn area.The dataset will have initial success metrics if the overall **MSA**(individual measures of sampling adequacy of each variable) $>$ **0.5**.This will be achieved by using **Factor Analysis** with **Kaiser-Meyer-Okin(KMO) Test** analysis using R package. Variables with a KMO-MSA value greater than 0.5 is considered adequate, and greater than 0.8 is considered meritorious.This procedure will help eliminate uncorrelated factors before training data.
- The most important success metric is to see if this project could potentially benefit the State and U.S Department of Agriculture on reducing firefighting costs in the next decade.The current yearly cost is around $1.4 billion, and the projected cost in 2025 will rise to $1.8 billion.By successfully identifying attributes that are bigger factors of fire area, CAL-FIRE and other fire fighting and prevention program can develop effective strategies down to all the county levels and reduce the expenses.

Estimated Risks

- Wild fires are unpredictable and random, therefore the it varies from year to year. It will take many years to monitor the cost trend.It is difficult to see at this time whether the project will give success metrics.
- Some statisticians believe that a MSA score of 0.5 is mediocre and should be above 0.6.Therefore, this dataset might not give a strong predcition.
- The climate and geographical properties in Portgual is different from ones in Nothern America especially California. Hence, the associated variables such as *Rain*, *Relative Humidity*, *Temperature* will have different values and the result will differ and might not speak for California.

Deployment

Users of this model/analysis

The users of the model/analysis are for Chair of Board of Fire and Forests Protection - Dr.J. Keith Gilless, and CAL-FIRE Director -Ken Pimlott, Office of The State Fire Marshall, UC Berkeley Fire Prevention Engineering program, Research Scientists, Environmentalists, related fire training and education programs, federal and state legislatures.

How the Model/analysis will be used

Based on this model analysis, **CAL-FIRE** can collect data and formulate the dataset using the same identified attributes associated with fire area size and make predictions of the potentially affected area across different months and days.

Risks associated with Deployment

Some attributes in the dataset such as Forest Fire Weather Index(FWI) is based on the Canadian system for rating fire danger, thus it may not apply for U.S system.

The other risk is that the machine learning models results are inconclusive.

Proposed Project Timeline (continuously updated based on the progress)

1. **Initial Analysis** complete by **4/15**. This includes:
    - Create a RMarkdown document named project-performance.Rmd
    - Peform Data Cleaning/Transformations and Exploratory Data Analysis using ggplot2
    - Factor Analysis using KMO Test
2. **Simple Analysis** complete by **4/24**. This includes:
    - Create an asset list(ASSETS.md) documenting progress and findings.
    - Plot histogram & predictors.
    - Finalize **EDA** with Correlation Matrix and Factor Loadings.
    - Evalulate and interpretate linear model.
    - Buld and evaluate a tree model
3. **Model Development** complete by **5/3** This includes:
    - use different ML techniques and compare results
    - Linear Regresson
    - Support Vector Machines (SVM)

    - normalize the datset
  - other methods (pending)
  - Assess the accuracy of the models.
  - Document the work flow for fitting a SVM, Regression for classification in R.
4. **Model Deployment** complete by **5/24**
5. **Model Management** complete by **5/30**