ILLNESS Score: A Multi-Factor Machine Learning Model for Wildfire Risk and PSPS Impact Assessment

Gloria Kao gkao@ucsd.edu
Shentong Li shl091@ucsd.edu
Neil Sharma n9sharma@ucsd.edu
Phi Nguyen pnguyen@sdge.com

Abstract

Wildfires in Southern California are a significant environmental threat, particularly in areas serviced by San Diego Gas & Electric (SDG&E). To mitigate this risk, SDG&E relies on Public Safety Power Shutoffs (PSPS), traditionally determined by windspeed thresholds. However, this approach can be overly simplistic, sometimes leading to unnecessary power disruptions. In this project, we propose the development of the ILLNESS score, a more comprehensive metric that considers additional environmental and social factors, such as temperature, humidity, and community impact, to determine the need for PSPS more effectively. The ILLNESS score—Insight on Life, Living, Nature, Energy, Service, and Season—aims to provide a balanced, data-driven approach for evaluating wildfire risks. We will develop a map-based web application to visualize ILLNESS scores as a heatmap for SDG&E's internal use and public transparency. This tool will empower communities with understandable, data-backed information about potential power shutoffs, improving both public safety and communication.

1 Introduction

1.1 Broad Problem Statement

Currently, SDG&E employs Public Safety Power Shutoffs (PSPS) as a preventive measure to mitigate wildfire risks. The decision to activate a PSPS is primarily based on windspeed thresholds, which can sometimes be overly simplistic, resulting in unnecessary power disruptions. Given the complexity of wildfire risks, which are influenced by a range of environmental and social factors, there is a need for a more comprehensive, data-driven approach to PSPS decision-making. This project aims to develop the ILLNESS score—a novel metric that incorporates multiple dimensions of risk, including weather conditions, community impact, and infrastructure vulnerability. By enhancing the accuracy and transparency of PSPS activation, the ILLNESS score seeks to improve wildfire mitigation strategies, reduce unnecessary power disruptions, and increase public trust in safety measures.

1.2 Narrow-down Problem Statement

The existing approach used by SDG&E to evaluate wildfire risks is heavily reliant on meteorological data, particularly windspeed thresholds, which may not fully capture the multifaceted nature of wildfire risk. The ILLNESS score will assist SDG&E staff in making informed decisions regarding PSPS activation. The score, which is calculated using various weighted environmental and social factors, will provide a more nuanced assessment of wildfire risks compared to the existing meteorology-driven approach. By presenting the ILLNESS score as a single, easy-to-interpret metric, it will help internal teams evaluate the likelihood of wildfire incidents with an information-dense value and determine whether a power shutoff is necessary. Additionally, internal staff will have access to detailed insights and analyses report of the individual factors that contribute to the score, allowing for more targeted and efficient wildfire mitigation strategies.

1.3 Literature Review

1.3.1 Electricity Distribution in Southern California

Southern California is served by a complex electricity distribution network managed by investor-owned utilities like SDG&E. Covering a service area of approximately 4,100 square miles and supplying electricity to 3.7 million people, SDG&E operates extensive infrastructure that includes both high-voltage transmission lines and local distribution systems (Electricity Forum (2024)). The distribution network must adapt to various challenges, including the risk of conductor breaks and faults due to external factors like vehicle collisions, weather events, and vegetation interference. In recent years, climate change has intensified these risks, leading to higher frequencies of wildfires and more unpredictable weather patterns, including the notorious Santa Ana winds, which exacerbate fire dangers.

To safeguard public safety and infrastructure, SDG&E has incorporated advanced technologies such as phasor measurement units (PMUs) that capture high-resolution data on phase angles, voltages, and currents across the grid (Schweitzer Engineering Laboratories (2024)). This real-time data, synchronized across multiple points, allows for rapid detection and response to anomalies like conductor breaks, reducing the time it takes to cut power in case of a fault. SDG&E's Falling Conductor Protection (FCP) solution, developed in collaboration with Schweitzer Engineering Laboratories (SEL), is one example of such innovation. The FCP system uses PMUs to detect potential conductor breaks and isolate affected circuit sections within milliseconds, preventing potential ignition sources from reaching the ground and minimizing the risk of wildfire.

1.3.2 Previous Actions on Wildfire Mitigation

California's utilities have taken proactive measures in recent years to mitigate the risks posed by wildfires. The 2019 California Public Safety Power Shutoffs (PSPS) highlighted the severity of the wildfire threat (Wikipedia contributors (2024)), as utilities like PG&E,

Figure 1: The Bond Fire, driven by high winds, burns near Silverado Canyon in 2020

SCE, and SDG&E preemptively cut power to millions of residents to prevent fires from igniting due to downed power lines. These widespread PSPS events, while effective in reducing fire risks, drew public backlash due to their impact on communities, especially those dependent on electricity for medical needs and critical services. The challenges faced during these events underscored the need for more sophisticated solutions beyond broad shutoffs.

SDG&E, in particular, has implemented a series of targeted programs to address wildfire risks. The utility's Fire Threat Reduction model, developed following major fire events, uses fire history, ignition data, and environmental conditions to map high-risk areas (San Diego Gas Electric (2016)). Additionally, SDG&E's Fire Potential Index (FPI) and the Santa Ana Wildfire Threat Index (SAWTI) provide predictive insights into fire risks based on wind conditions, fuel moisture, and temperature. These indices help SDG&E prioritize de-energizing high-risk circuits and guide operational decisions on power restoration during dangerous conditions. Furthermore, SDG&E has adopted advanced vegetation management practices and improved system protection settings to minimize fire hazards associated with its distribution lines. These efforts illustrate the utility's commitment to reducing wildfire risks through data-driven approaches and continuous innovation.

2 Methodology

2.1 Weather Data Analysis

We consider two primary sources of weather data: forecast data and observational data collected from SDG&E's weather stations. The forecast data is critical for training our model, particularly because they are likely to be used as inputs in various applications in case this project wants to be altered. In contrast, observational data, which are only available post-event, offer critical insights into the differences between predicted and actual weather outcome.

2.1.1 Data Structure and Processing

The forecast data were divided into two datasets: one containing only weather information and another incorporating wildfire risk statistics. The data was originally collected on an hourly basis, resulting in a dataset of impractical size for training with our available computational resources. To address this challenge, we aggregated the data by longitude-latitude coordinates and computed the daily average of the features, thereby reducing the data volume by a factor of 24 while preserving key patterns.

Although our weather model utilized more than a dozen features, we concentrated on humidity, temperature, and wind speed as the primary drivers of wildfire activity by employing an attention mechanism. Our target variable was defined as a composite index derived from fire-related parameters, including the ignition component, fire intensity level, and forward rate of spread.

2.1.2 Weather Model

The concept of predicting wildfire risk based on weather data is extremely complex, and therefore we needed a sophisticated model to capture the nonlinear relationships inherent in the data. To address these challenges, we used a multi-layer perceptron model enhanced with an attention layer. This architecture allowed the model to selectively focus on the most relevant meteorological features, thereby improving its predictive performance and robustness in assessing wildfire risk.

The dataset was split into training (80%) and testing (20%) sets. The model was evaluated using:

- **Mean Absolute Error (MAE):** 0.8361
- **R-squared (R^2) Score:** 0.7284

Given that the output is a wildfire risk score on a scale from 0 (no risk) to 10 (high risk), an MAE of 0.8361 and an R^2 of 0.7284 indicate that the model performs reasonably well. However, there remains room for further improvements in predictive accuracy.

2.2 Conductor Data Analysis

The conductor data includes a large span of data, including the length, structure material, upstream/downstream spans, work order dates, and more. We focus on the structure material, age, type, wire risk, and days since the last work order of each conductor span, then correlate them to a wildfire risk value. From our exploratory data analysis, we can infer that steel structures and overhead powerlines correlate to the highest public safety power shutoff (PSPS) probability. Some of these variables are categorical, so we one-hot encoded them, and then applied a machine learning algorithm.

2.2.1 Machines Learning Model

A Logistic Regression model would have been suitable for categorical data, but we want our model to calculate a scalable performance score, not predict a binary category. Thus, we tested Linear Regression and Random Forest models, which showed similar results in terms of accuracy. However, Random Forest has a slightly lower error, as recorded below:

- **Mean Absolute Error (MAE):** 1.497475
- **Mean Squared Error (MSE):** 9.986280
- **R-squared (R^2) Score:** 0.526328

Generally, a R^2 Score above 0.7 would be considered more significant, so our model still has some improvements to be done in the next few weeks.

2.3 Living Data (Customers)

The living component of our analysis focuses on customer-related factors.

Since the number of customer-related variables is relatively small compared to vegetation and geographic factors, we integrate this dataset into the Vegetation & Geographic Analysis model. This ensures a more comprehensive assessment of wildfire risk by incorporating human impact factors alongside environmental predictors.

2.4 Vegetation and Geographic Data Analysis

To assess the impact of vegetation and geographic factors on wildfire risk, we analyzed data from the Vegetation Resources Inventory (VRI), geospatial elevation data, and customer distribution. Our goal is to quantify how these variables contribute to the likelihood of Public Safety Power Shutoffs (PSPS) and wildfire risk.

2.4.1 Data Processing

We integrated multiple datasets, including:

- **VRI Data:** Categorized into high (H), medium (M), and low (L) risk levels, representing vegetation density and wildfire risk potential.
- **Geographic Data:** Elevation, latitude, and longitude extracted from GIS records.
- **Customer Data:** Number of total customers, life-support-dependent customers, urgent care users, and medical-certified customers in each region.
- **Tree Density Data:** Number of trees at risk of striking electrical infrastructure.

These datasets were merged using a spatial join to associate vegetation risk levels and geographic features with weather station records.

2.4.2 Machine Learning Model

We employed a **Random Forest Regressor** to model the relationship between vegetation risk, geographic attributes, and wildfire probability. The Random Forest model was selected due to its ability to handle non-linear relationships, its robustness to outliers, and its capacity to determine feature importance.

Model Training and Evaluation: The dataset was split into training (80%) and testing (20%) sets. The model was evaluated using:

- **Mean Absolute Error (MAE):** 0.0835
- **Mean Squared Error (MSE):** 0.1648
- **R-squared (R^2) Score:** 0.9930

indicating high accuracy and strong predictive performance. For comparison, a Linear Regression model was also tested, but it performed significantly worse, achieving an R^2 score of only 0.4631.

3 Result

4 Discussion

5 Appendix

The primary objective of this project is to develop the ILLNESS score, an acronym for Impact on Life, Living, Nature, Energy, Service, and Season. Using these 6 variables, we aim at creating a composite scoring system to analyze risk of Public Safety Power Shutoff (PSPS) events as a proactive measure at combating wildfire risk.

The score's primary goal is to enhance SDG&E's decision making process regarding wether or not to PSPS an area, improving the robustness and accuracy of these critical decisions. This scoring system may also be used for mapping PSPS risk over SDG&E's territory, providing transparency to the public, government entities, and stakeholders.

6 Contributions Statement

Neil Sharma: 34% (code + report writing)

Gloria Kao: 33% (code + report writing)

Shentong Li: 33% (code + report writing)

References

Electricity Forum. 2024. "Electrical Distribution Systems Explained." [Link]

San Diego Gas Electric. 2016. "OIR Fire Threat Mapping CPUC Presentation." [Link]

Schweitzer Engineering Laboratories. 2024. "Mitigating Wildfire Risk in Southern California." [Link]

Wikipedia contributors. 2024. "2019 California Power Shutoffs." *Wikipedia, The Free Encyclopedia*. [Link]