Predictive Earthquake Damage Modeling for Natural Gas Distribution Infrastructure

by

Steven B. Link

B.S. Systems Engineering, United States Naval Academy, 2010

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering in partial fulfillment of the requirements for the degrees of

Master of Business Administration
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature of Author ..... Signature redacted
MIT Sloan School of Management
Department of Mechanical Engineering

Certified by ..... Signature redacted .....
Georgia Perakis, Thesis Supervisor
William F. Pounds Professor of Management Science
MIT Sloan School of Management

Certified by ..... Signature redacted .....
Saurabh Amin, Thesis Supervisor
Robert N. Noyce Career Development Associate Professor
Department of Civil and Environmental Engineering

Certified by .. Signature redacted .....
Konstantin Turitsyn, Thesis Reader
Associate Professor
Department of Mechanical Engineering

Approved by ..... Signature redacted .....
Rohan Abeyaratne, Mechanical Engineering Graduate Committee Chair
Department of Mechanical Engineering

Approved by ..... Signature redacted .....
Maura Herson, Director of MBA Program
MIT Sloan School of Management

THIS PAGE INTENTIONALLY LEFT BLANK

Predictive Earthquake Damage Modeling for Natural Gas Distribution Infrastructure

by

Steven B. Link

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering on May 11, 2018 in partial fulfillment of the requirements for the Degrees of Master of Business Administration and Master of Science in Mechanical Engineering

Abstract

The Pacific Gas and Electric Company (PG&E) operates and maintains 48,000 miles of natural gas pipeline, serving over 4.3 million customer accounts. Along with water, electric power, and transportation services, these lifelines serve critical functions throughout multiple communities.

Considering PG&E provides services in both densely populated and seismically active areas, the organization has invested extensively in modeling technology to help estimate resource needs and develop resiliency plans in the event of an earthquake. This thesis aimed to develop a damage prediction model to improve emergency response time and restoration efficiency.

The machine-learning based model built upon currently used predictive algorithms, while adding features necessary to account for distribution branch lines and above-ground meter sets. Research and analysis showed factors beyond ground-motion prediction equations could be used to estimate pipeline damage and were consequently included. Furthermore, the model incorporated real-time data acquired throughout repair and restoration efforts in order to improve the predictive performance. Historical incidents were examined in the data aggregation phase in order to develop the training set.

For this paper, damage was defined as the number of leaks predicted in a given plat, as defined by PG&E's mapping services. Leaks were categorized in three separate bins, ranging from 0 leaks, 1 to 5 leaks, and 6 or greater leaks. Multiple classification algorithms were chosen and evaluated against a custom scoring metric designed to discriminate and penalize false negatives. The best results were achieved using a series of five logistic regression algorithms, executed at 2, 4, 8, 12 and 24 hours following event occurrence.

Results were designed to accompany currently used seismic hazard reports in a ranked table, displaying areas with the highest to lowest probability of experiencing damage. An additional web application was designed to query specific plats for prediction results.

Thesis Supervisor: Saurabh Amin
Title: Robert N. Noyce Career Development Associate Professor, Department of Civil and Environmental Engineering

Thesis Supervisor: Georgia Perakis
Title: William F. Pounds Professor of Management Science, MIT Sloan School of Management

Thesis Reader: Konstantin Turitsyn
Title: Associate Professor, Department of Mechanical Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

I would like to thank my advisors Prof. Georgia Perakis and Prof. Saurabh Amin. Their guidance and dedication over the last year has been truly invaluable. I would not have been able to accomplish this project without their knowledge and support.

In addition to my advisors, I would also like to thank Mathieu Dahan (PhD candidate MIT Dept. Civil and Environmental Engineering) for his expertise. His approachability and skill sets were critical in maintaining progress.

Thirdly, the entire Pacific Gas and Electric company made this project a special experience. They consistently provided the necessary support, resources, and time needed to help me accomplish the project. Beyond their knowledge and professionalism, I always felt welcomed and part of the team. In particular, Bryan Hennessy and George Gaebler were exceptional before, throughout, and after my time on-site came to a close.

Lastly, I would like to thank my entire family for their love and support over the last two years at MIT. Specifically, I owe a lot to my wife Esmeralda who has been nothing short of amazing through it all.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

1 Introduction to the Earthquake Damage Prediction Project at PG&E

1.1 Company Overview

Incorporated in 1905, Pacific Gas and Electric (PG&E) Company provides natural gas and electric services to over 16 million people throughout the state of California. With a 70,000-square-mile service territory, roughly two thirds the size of the state, PG&E relies on an extensive network of infrastructure to safely carry out the transmission and delivery of energy [1]. Regulated utilities in the state of California do not own production facilities, and instead receive 91% of their natural gas from a series of interstate pipelines. These lines originate from basins in Canada, the Rocky Mountains, and Texas, before reaching California and entering local storage, transmission, and distribution systems [2].

PG&E provides 2.6 BCF of natural gas on a daily basis to a variety of customers ranging from residential to large-scale industrial consumers and electric generators. In order to accomplish this, they utilize 7,000 miles of transmission pipeline operating between 600 and 60 pounds per square inch gauge (psig). A series of compressors are used to maintain the appropriate pressure while moving gas through the system. The distribution system comprises of 48,000 miles of pipeline, along with a series of regulators and meter stations that enable natural gas to reach the final customer. Beyond enabling the use of natural gas at the consumer level, the transition point between transmission and distribution systems (60 to 0.25 psig) also includes the addition of ethyl mercaptan. This odorant provides a warning system for gas leaks, as natural gas on its own is odorless [3, 4].

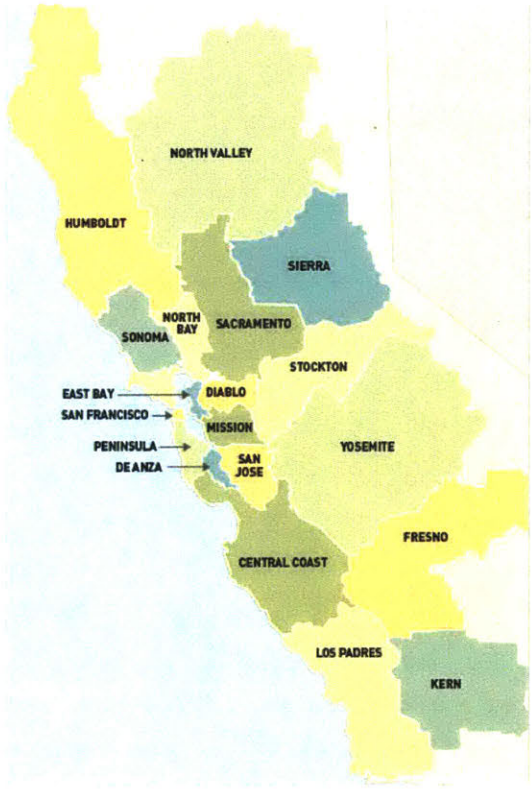


Figure 1: PG&E service territory labeled by division

1.2 Current Post-Earthquake Response

The responsibilities incurred as a pipeline operator are heavily regulated at the Federal and State level. Geographically, however, PG&E faces unique challenges beyond what is typically seen by an investor owned utility. In 2002, the United States Geological Survey (USGS) formed a series of working groups to specifically examine the likelihood of a major earthquake occurring in the San Francisco Bay Region (SFBR). They concluded that there is a 62% probability of an earthquake with a moment magnitude (M) greater than 6.7 occurring before 2031. To give context on the associated damage, the report cites two previous earthquakes of M6.7 (1994 Northridge, CA) and M6.9 (1995 Kobe, Japan) which resulted in $20B and $147B in damages respectively [5].

PG&E utilizes an internal document known as the Gas Emergency Response Plan (GERP), which is an annex to the Company Emergency Response Plan (CERP). This document outlines detailed emergency response guidance for personnel to safely and efficiently respond to gas system emergencies and restore operations to their normal state. Developed and maintained through Gas Emergency Preparedness (GEP), the GERP provides an excellent overview and is augmented through company earthquake-specific documentation such as the Asset Knowledge and Integrity Management Earthquake "Playbook."

PG&E maintains five emergency levels which dictate the activation of various emergency procedures. The levels are numbered and range from "routine" to "catastrophic" and are driven based on the severity of the event, number of affected customers, injuries/damages, and other factors that are listed in detail in the GERP. An earthquake with a magnitude of 6.0 or greater will drive the organization to stand-up the Gas Emergency Center (GEC) and Operations Emergency Center (OEC). The GEC is located in San Ramon and OECs are formed at the divisional level. PG&E gas operations are split among 15 divisions as seen in Fig. 1. Based on the location of the damage and severity of the event, multiple OECs can be established. If needed, PG&E will also establish the Emergency Operations Center (EOC) in the General Office in San Francisco. This center will control the management of both gas and electric assets during an emergency.

In regards to the distribution system, the response following an earthquake can be viewed as two largely independent, but parallel processes. They include the Distribution Integrity Management Program (DIMP) team, which provides guidance on where to survey for leaks, and the Gas Service Representative (GSR) team, which immediately responds to emergencies tasked through Dispatch. Fig. 2 is an illustrative example of how information flows among the relevant parties and Appendix A provides a detailed narrative regarding to roles and responsibilities of the aforementioned teams. Information regarding the Dynamic Automated Seismic Hazard (DASH) model currently used to aid emergency responders is also presented in Appendix A and Appendix B.

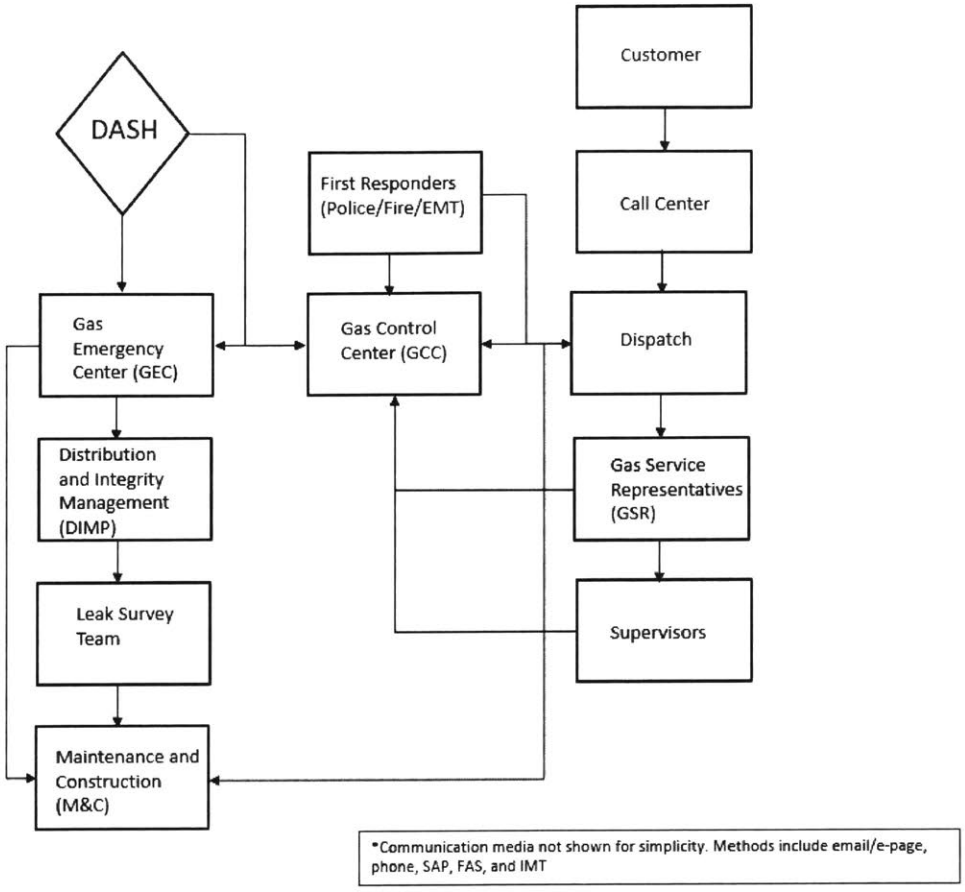


Figure 2: Information flow following an earthquake (flow chart non-exhaustive and relates only to applicable aspects of this specific project)

1.3 Thesis Motivation

In order to provide the safest and most reliable service, PG&E has invested extensively in modeling technology to help estimate resource needs and develop resiliency plans. Seismic events are unpredictable, which makes timely data acquisition and analysis a critical component in responding quickly and effectively after a large event.

Gas Control aims to be the front line for public and employee safety and system reliability. Damage to natural gas infrastructure can be catastrophic, resulting in inhalation, fire, and even explosive hazards. Additionally, restoration of electrical systems cannot be completed until the area has been swept to ensure no gas leaks are present. Effective identification and efficient restoration is critical to ensure people are not displaced for lack of basic needs.

To ensure this is accomplished, Gas Operations has the strategy to "transform data into intelligence to operate predictively and proactively in order to identify and mitigate risks in real time" (PG&E Gas Operations vision statement). The proposed improvements to the model directly align with this strategy.

Additionally, important emergency response and restoration efforts are being driven through models originally designed for water pipes (discussed further in Section 3.2) [3]. They represent an adequate baseline but there is significant room for improvement as more data becomes available. Ample time has

passed since the last earthquake with a moment magnitude greater than 6.0 occurred in Northern California (American Canyon 2014), meaning there is now the opportunity to examine the results and develop hypotheses for new prediction methods.

Also, the current model only takes into consideration damage to below-ground, main lines. Assets such as service lines and meter sets have not been considered. PG&E has made great efforts to map all of their assets in order to have them digitally available. With the increase in information and ease of access, it is reasonable to suggest that a model can be effectively created to take such things into consideration.

Lastly, financial concerns provide incentives for improving emergency response efforts. Failure to accurately account for assets damaged in the event of an earthquake can lead to future issues and possible incidents that fall within company control. Beyond the personal injury and loss of life, potential property destruction and gas loss related to damaged pipelines cannot be ignored. Since 1997 gas utilities in the state of California have lost $662.5M as a result of faulty pipelines [6], which is a considerable sum but only a fraction of the potential multi-million-dollar fines and lawsuits possible, as seen from the 2010 San Bruno explosion [7].

1.4 Thesis Hypotheses

The literature findings, interviews with PG&E employees, and careful review of company, city, and state post-earthquake after-action reports helped create a holistic understanding of the elements involved in lifeline damage during a seismic event. Based on these findings, we hypothesized that additional, unimplemented factors could improve the performance and scope of the existing damage prediction model. Mainly, currently available data could be used more efficiently. Secondly, by incorporating real-time data from customers and first responders, the predictive power of the model could significantly increase. New data, available throughout the life an event, provides invaluable information and should be included to enhance model effectiveness. Such analytically driven operations can then be used to optimally allocate resources for restoration efforts.

1.5 Literature Review

1.5.1 Introduction

Building robust infrastructure to handle the stressors induced during earthquakes is not a new concept, but there are many assets that remain susceptible to ground shaking and rupture. Actions can be taken to retrofit facilities, and rigid pipeline material such as cast iron can be replaced with more flexible plastic and steel. Unfortunately, earthquakes tend to inflict damage in new and unpredictable ways. Before examining potential methods to better predict damage to natural gas pipelines, it was critical to review the most current ground motion prediction equations and fragility models, along with empirical data from past events within the PG&E service territory. Following this, predictive analytic concepts and machine learning methods were examined as tools for improving emergency response operations in the event of an earthquake.

1.5.2 Damage Prediction Tools and Models

Government, academic institutions, and private organizations routinely use a Federal Emergency Management Agency (FEMA) developed software known as Hazards U.S – Multi-Hazard (HAZUS-MH). Extensive studies and evaluations of the software's effectiveness have been conducted, indicating its prominent use as the standard risk assessment and loss estimation software. The tool is designed to examine the effects of floods, earthquakes, hurricanes, and tsunamis, while providing insight on the

physical, economic, and social impacts of the disaster. Of note, Lewis County, Washington and Vancouver, British Columbia (using software adapted for international use) have used the modeling software in conjunction with USGS probabilistic ground motion data to predict possible earthquake damage. These predictions have since been used to develop emergency response plans and retrofit vulnerable infrastructure to withstand ground motion and failure effects. In 2014 (American Canyon earthquake) the city of Napa used the software to cross-reference their damage inspections with HAZUS-MH predicted results to validate their findings [8].

FEMA publishes user and technical manuals for the software, detailing the methodology used to predict damage, personal injury and death, and the resulting economic losses. Sources and expert opinion regarding asset and building Inventories, demographics, and economic parameters are listed to provide additional information to the user to understand the final predictions and recognize the potential uncertainty involved. Section 8.4 of the Multi-Hazard Loss Estimation Methodology (Earthquake Model) Technical Manual [9] details damage functions (fragility models) for buried pipeline. The functions used are the same as those designed for oil pipelines and potable water pipelines. The extent of the damage is described through a repair rate, which primarily results from the buckling of the pipe wall. This means that peak ground velocity (PGV) and flexibility of pipe material and welds have been designated as major indicators for predicting leaks. Specifically, polyethylene pipes with fused joints, ductile iron and steel pipes with ball and spigot or arc welded joints are considered the least vulnerable to damage from ground shaking [10]. Repair rate equations have been listed in Appendix C. Peak ground disturbances (PGD) such as a major ground rupture often result in the breaking of the pipe, which is more difficult to mitigate through material selection. Instead minimizing pipe placement over major faults has been employed, but this is more applicable to transmission lines and not further explored in this document.

PG&E also uses fragility models based on water systems while accounting for pressurization differences [3,11]. Their repair rate equations are listed in Appendix C, along with scaling factors designed to reflect current pipeline materials. Of note, the effectiveness of ductile pipes in withstanding ground motion can be negated through corrosion, and as a result, soil corrosiveness factors have also been included in company damage prediction calculations.

1.5.3 Historical Events

All studies have not been listed, but several key resources are worth mentioning in order to best understand the data collection phase of the project. Major events within PG&E's service territory were examined in order to accurately assess the risks associated to their infrastructure and customer base. The last two earthquakes in the region to have warranted activation of the Emergency Operations Center (EOC) occurred in 1989 (Loma Prieta earthquake) and 2014 (American Canyon earthquake). The San Francisco earthquake of 1906 was studied less extensively based on the drastic changes in building and pipeline construction, but it still provided important historical records. Regarding the 1906 earthquake, Douglas G. Honegger's report compares leak data with soil deposits and peak ground acceleration [12]. Numerous internal company reports were also provided for this thesis, of which many are publicly available. Of note, Steven H. Phillips and J. Kris Virostek compiled an extensive after-action report for the Loma Prieta earthquake [13], which provides a critical review of the event. Additional resources from the California Seismic Safety Commission [14] highlight additional lessons learned. Finally, Hope Seligson and MMI Engineering conducted several in-depth studies regarding the effectiveness of the currently used earthquake damage prediction model at PG&E based on the 2014 American Canyon Earthquake, along with methods for expanding model coverage [3, 15, 16].

1.5.4 Machine Learning Applications

Applying predictive analytics and machine learning principals in an effort to improve major utility operations has been on-going for some time. Poisson regression algorithms were initially used to identify most-likely points of failure [17] for utilities during weather events, and PG&E continues to invest in new data driven improvement projects designed to drive everyday operations such as vegetation clearance schedules [18]. Many times, these applications are utilized by electric power providers, which can take advantage of rapidly developing, and numerous observation points associated with above-ground lines stretching vast distances.

Less frequent occurrences, however, are also being studied through predictive analytics. Landslides in particular are often discussed and multiple studies have been conducted to examine predictive spatial features that may provide valuable warnings regarding susceptible areas. Decision trees and support vector machines (SVMs) have been explored with successful results [19, 20] along with various other algorithms using Geographical Information Systems (GIS) [21]. Beyond this, traditional machine learning applications for the use of aiding earthquake recovery are often focused on remote sensing image analysis to assess damages [22].

Additional statistical methods for natural disaster early warning have been discussed in the realm of sentiment analysis and natural language processing in filtering Twitter and social media posts regarding earthquakes and wild fires [23, 24].

1.6 Thesis Contribution and Outline

This thesis makes several important contributions, in which the methodology and results are described throughout the next several chapters. Model benefits described in this work include:

1) Expanding the currently used model to include branch lines and above-ground assets

Taking into consideration these lines and above-ground assets (such as meter sets) greatly expands the robustness and impact of the model. Previous methodologies perform well on below ground transmission lines but are ill-equipped to make predictions on such components of the distribution system. This a critical and novel addition in PG&Es tool set to effectively and safely restore operations following an earthquake.

2) Improving predictive power through the incorporation of additional features and real-time data

This is the first time that machine learning techniques have been used to predict gas line damage following an earthquake. Previous models have solely relied on ground motion prediction equations based on empirical evidence and expert opinion. This model utilizes such equations, along with crowd-sourced information to actively train during an event in order to improve predictive power.

3) Eliminating manual procedures by automatically compiling data from multiple internal and external sources

Immediately following an event, information is compiled from multiple internal and external datasets in order to begin implementing the model. Prior to this, multiple departments had to rely on informal information transfer and latency before having the information required to make the most informed decision.

4) Establishing a framework for future model development

The tool and methodology are data-driven and will improve following each event. While the results of this thesis are promising, they are largely based on the data acquired from a single event. As soon as new information is acquired, whether from historical events covered by other utilities, or in real-time immediately following the next emergency response, the described model framework can be utilized to make damage predictions among affected plats.

In order to reach these benefits, Chapter 2 of this thesis presents the data sources used and the transformation techniques required to create a mineable dataset. Information from internal company servers, Federal Emergency Management Agency (FEMA) simulations, and United States Geological Survey data had to be compiled and transformed. GIS modeling techniques then improve data interpretation and provide meaningful outputs based on PG&Es current operating procedures. Lastly, the manual compilation and analysis is automated through Python script in order to integrate with other PG&E systems in a timely and hands-off manner.

Chapter 3 shows the methodology used for developing the predictive model, highlighting various machine learning techniques to optimize performance. Model evolution is described through the development of both a static and dynamic, iterative model that updates and improves throughout the duration of an event. Hypothesized predictor variables are consequently added and removed as individual algorithms and their respective hyperparameters are tuned against a custom performance metric. Finally, the results are reviewed and compared for final model selection.

Chapter 4 examines the model output and describes how it can be effectively implemented within PG&E's Gas Emergency Center (GEC). Proposed uses include providing leak probability predictions to the Distribution Integrity Management Program (DIMP) to drive the deployment of Leak Survey crews and to use aggregated predicted leaks on a divisional level to feed the Maintenance and Construction (M&C) resource optimization model.

This thesis concludes with alternative model use recommendations and areas of future work in chapter 5. Additional algorithms that were not discussed in this document and unexplored datasets could provide benefit to the organization in the future if additional time and resources can be afforded.

2 Data Collection and Resources

2.1 Data Sources

Prior to model development we had to construct a database from PG&E asset information and historical records from previous earthquakes. These records include reports from industry, government, and academia, focused on documenting the impact of specific earthquakes and reviewing measures that can be taken in the future to minimize damage and loss of life. The below paragraphs outline specific data sources used and the reasoning behind their inclusion.

Predicting the number of leaks that may occur in each plat requires an understanding of the assets present in the area and their function. Based on empirical data from previous earthquakes, the focus of this thesis is on the distribution system, as it is more vulnerable to the effects of an earthquake and largely present in heavily populated areas. The distribution system will refer to lines downstream of the regulator stations and less than 60 psig. These lines form a below-ground network, often running alongside other utilities such as electric, water, and telecommunications, in order to deliver natural gas from the transmission system to the customer. Smaller diameter service lines then connect the distribution main lines to a gas

meter at the service tee through branch or service lines. This point represents the boundary between utility and customer owned facilities. See Fig. 3 [25] for an illustration of the distribution system components.

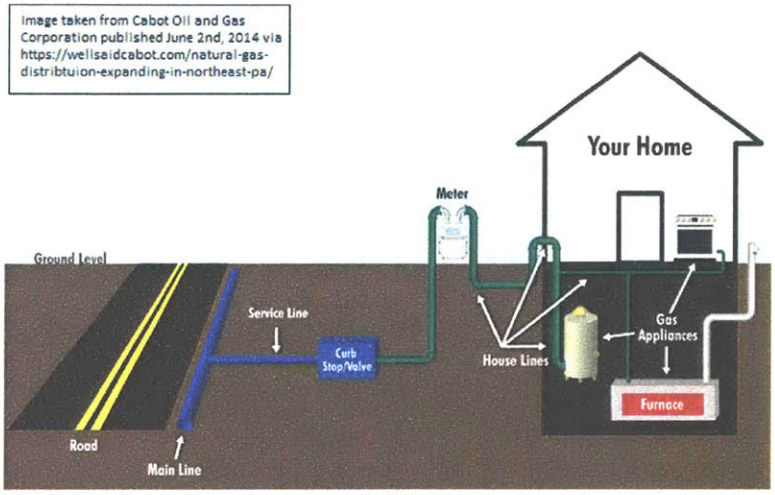


Figure 3: Illustration of the gas distribution system and terminology

For our dataset, we were unable to attain the length of the service lines. Based on the number of accounts and size of the service territory, documenting this number is still an on-going process. To mitigate this missing set, we were able to identify the length and material of all main lines, and the number of service taps within each plat. With these data we could get a reasonable sense of the assets in the area.

Ground displacements, which will be discussed later in the chapter, can have a significantly different impact on the system depending on the type of pipe present. Sufficient flexibility or the ability to force soil movement around the pipe is dependent on the material composition and construction techniques. This required us to identify and differentiate between welded steel pipelines and those constructed of medium to high-density polyethylene [14]. Our final dataset includes the breakdown of pipeline material within each individual plat.

Thirdly, Field Automation System (FAS) data provided a record of every call PG&E received, from both customers and first responders (Fire, Police, EMT, etc.) during the 2014 American Canyon earthquake. In the event of an emergency, customers notify the PG&E Call Center if they smell gas (or think they smell gas), are concerned and want a representative to check their appliances or have manually shut-off the gas and need a qualified technician to restore the service. The customer level of knowledge regarding the distribution system and their own gas appliances varies widely, resulting in a high ratio of calls to actual confirmed leaks. Therefore, calls alone are not a consistent proxy for determining gas leaks, but can often occur in response to excessive shaking (different building structures may react more violently with lower peak ground acceleration (PGA) values based on the type of construction and soil composition on which they are built) and other factors that may be helpful in predicting gas leaks. Therefore, the call data was collected and included within the dataset of potential predictor variables.

Census data was also pulled from the ArcGIS "living atlas," an extensive database of global geographic information from Esri. The data came from a 2012 census, which was assumed to be an accurate representation of the population in the Napa Valley in August of 2014. Population was split into tracts and later utilized to normalize additional data such as customer calls. This was done to avoid assigning higher

weights to the call values from more populous regions, where multiple calls could be in reference to a single incident.

Red-tagged building data from the Earthquake Engineering Research Institute (EERI) Napa Valley Reconnaissance Survey and predicted building damage from the HAZUS-MH software was also collected. These data were used to evaluate the hypothesis that building damage could act as a predictive feature in determining the location of gas leaks. The reasoning behind using both predictive values and historical records is listed in Section 3.

Historic leak rates were further identified as a potential predictor for leaks during an earthquake. Aside from material strength, things like corrosion and age of pipe can cause performance degradation. Without having access to the age of each pipe segment or the direct corrosive effects, we decided to look at previous survey results to account for such factors. This allowed us to identify plats that were more likely to contain pipes with minor, existing leaks or those that are more susceptible to future leaks.

It should be noted, any large natural gas distribution system may contain several hundred leaks of various magnitudes at any given time. Plats are surveyed periodically to identify and fix such leaks and are graded based on their severity and required repairs. A Grade-1 leak is determined to be hazardous to the public and must be repaired immediately. These would include events such as construction equipment rupturing a line (dig-ins) to non-hazardous leaks of any size, but that occur within 5 feet of a residential building. Grade-2 leaks require a scheduled repair within 15 months and periodic surveillance based on the condition they may become hazardous over time. Grade-3 leaks are considered non-hazardous with the reasonable expectation they will remain that way. Regardless, they still require re-inspection every 15 months [8]. Identifying plats with previous leaks (both in quantity and severity) could be used to improve model prediction power after an earthquake. This hypothesis was also consistent with an internal PG&E report that identified previous leak survey timing as a "better predictor of leak rate post-earthquake than peak ground acceleration" [26]. With this in mind we added the previous survey date, and the number and type of each historical leak for each plat in the dataset.

USGS and PG&E's Geosciences Department provided historical geological data for the project. For below ground assets, ground shaking does not typically inflict damage unless it induces significant soil failure, but permanent ground displacement is important to identify and understand when looking to predict damage. Through efforts between PG&E and InfraTerra Inc., we were able to assign landslide, liquefaction, and surface rupture susceptibility values to each specific plat. Details on the equations and accompanying factors as a function of PGA can be found in Appendix B. PG&E uses these factors to create Earthquake Prioritization (EP) values that determine which plats should be surveyed first following an earthquake. Scores are then normalized to 100 but can exceed that value if the PGA values are recorded above 0.5 g [27].

The direct PGA values from historic events have also been included in the dataset. As previously described, they correlate with permanent ground displacement, but can also be useful in predicting above ground damage. Masonry from chimneys and collapsed cripple walls have been documented as crushing and severing both meter sets and service tees, resulting in Grade-1 gas leaks. An internal memorandum [16] from MMI Engineering provides a detailed analysis on pre-1950 home construction in the San Francisco Bay area and the likelihood of collapsed cripple walls causing gas leaks on above-ground assets. The report further provides recommendations for incorporating this concept into the current system and what data sets would have to be compiled to make it possible. The recommendations are set for review and possible implementation, but to date no action has been taken.

During an actual event, these PGA values are received by PG&E directly though USGS ShakeMaps. They are inserted into internal algorithms to determine repair rates and make damage estimations. Building on this automatic transfer of information, we examined an additional feature of the USGS service (not previously used by PG&E), which involved crowdsourced information. The "Did You Feel It?" program [28] includes data collected from the general populace through an online survey. The survey answers are compiled and entered into an algorithm to determine the Community Decimal Intensity (CDI), which can be thought of as a separate intensity measurement. In the American Canyon earthquake, over 40,000 people responded to the survey, mostly within the first two hours. If addresses are included, USGS creates a geocoded map to indicate where the CDI values originated. The affected areas are then segregated into a grid (one-kilometer resolution) and are formed using UTM coordinate boundaries. For this dataset, the centroid of each affected area was chosen and matched with a corresponding plat. As mentioned in section 1.5.4, crowdsourcing information in the event of a natural disaster can be used to aid first responders.

Additionally, a historical set of reported leaks from the 2014 American Canyon earthquake was compiled. Without these data, the model would be unable to effectively learn. To accomplish this, we pulled data from SAP to account for the Leak Survey team reports, FAS to account for dispatched Gas Service Representative (GSR) reports, and the Gas Incident Management Tool (IMT) to account for major leaks that required responses from Maintenance and Construction (M&C) crews.

Through Leak Survey reports, we noted the teams were able to survey all plats with predicted damage within 6 days of the 2014 American Canyon Earthquake. Taking data from 24AUG2014 through 29AUG2014 in SAP, we were able to create a list of confirmed leaks, and the time in which they were discovered. Based on Leak Survey procedures, each documented leak already included the plat from which it was found.

It is important to note that leaks requiring minor repairs such as tightening, lubricating and adjusting (TLA leaks) were not included. These leaks exist throughout the service area and pose no safety hazard. Differentiating between pre-existing TLA leaks and those caused by the earthquake was not conducted in this analysis.

The FAS datasets served two important functions. Firstly, the data contained leak information that was not recorded in the Leak Survey reports. This is because GSRs and Dispatch utilize a separate system with their own requirements. These confirmed leaks were added to the SAP data discussed above. The second function (as previously described) relates to accessing customer phone call records in order to predict areas that may experience leaks.

Thirdly, data was collected from historical IMT records. IMT was utilized by dispatch to record incidents that required immediate action and met other reporting thresholds required by the California Public Utilities Commission (CPUC). These reports typically correspond to dig-ins, requests for qualified gas representatives by local fire and law enforcement, and other events that could result in significant damage, injury, or media attention. Notifications that result in IMT documentation most often originate through dispatch and will have an associated FAS record and work order as well. However, there are exceptions, and as such, IMT records were reviewed to complement the dataset compiled from FAS and Leak Survey reports. Using this methodology, we could effectively cross-reference all of the available resources to ensure we had the most accurate and comprehensive view of the leaks that occurred during the 2014 American Canyon earthquake.

2.2 Model Resolution

Based on available data limitations and unique damage profiles presented from each recorded earthquake, we felt it would be unwise to predict damage to individual main and service lines. For example, the 2014 American Canyon earthquake presented surface disruptions that were previously thought not to occur during earthquakes of that particular magnitude. Examples were discussed in an interview with members of the PG&E Geosciences Department (conducted March 1st, 2017) and are referenced throughout this work. Additionally, we did not have the asset data and positioning to attempt such levels of granularity. To provide meaningful results within our capabilities, we decided to examine individual plats. For this analysis in the Napa region, the average plat formed a rectangle approximately 400x600 meters. Plat size is dependent on the density of assets in the region and will therefore vary from region to region. The assumption is that plats are already used in the PG&E system for directing Leak Survey and other maintenance efforts, were small enough to provide meaningful direction to GSRs and other small crews and could be aggregated easily enough to input divisional leak totals to the resource optimization model.

Once limitations on granularity were set, it was important to transform the series of descriptive locational references into a singular system. For example, FAS data provided positions in terms of latitude and longitude (World Geodetic System 1984), IMT provided individual addresses, and census data came in the form of tracts. Addresses were geocoded and grouped along with geographic coordinates and stored as point data. Using ArcGIS software, these points were graphed along with a plat boundary layer and stored as polygon features. Intersecting the multiple layers allowed us to assign point values, such as confirmed leaks and customer calls to the plats in which they occurred.

While examining multiple layers stored as polygon features, we utilized the "tabulate intersection" tool, also available through ArcGIS software. In the case of determining population per plat, we examined the plat boundaries and census tract boundaries. The percentage of tract occupied by each plat was multiplied by the tract population in order to determine the population for each individual plat. For this calculation, tract and plat population were assumed to be uniformly distributed over their respective areas.

2.3 Summary of Data

A summary table of the original hypothesized independent variables is available in Appendix F. It is important to note that the table listed in Appendix F was an early representation and does not include imputed data or the elimination of independent variables through model iteration. From the available resources, a master dataset was constructed utilizing the plat names as indices. Plats were chosen based on the DASH model output, which provided a list of plats that could have possibly been affected by the earthquake. Using this list, independent variables were assigned from the variety of sources mentioned above. Finally, the dependent variable, represented by the number of leaks, was assigned to each corresponding plat. Preprocessing steps were taken based on the algorithm of choice. Chapter three will examine such steps and the decision points and methodology used throughout the evolution of the model.

3. Model Development

3.1 Overview

The methodology behind development can be summarized in five steps:

(1) Data and knowledge collection and analysis (described in Chapter 2)

(2) High-level solution design and feature selection

(3) Model type and algorithm selection

(4) Technical validation through training and testing

(5) Implementation

Chapter 3 focuses on steps (2) through (4) and details the specific methods and considerations used in accomplishing these tasks.

3.2 Incorporating Existing Architecture

Once an earthquake occurs, PG&E uses the DASH model to understand the event's impact to gas distribution assets. This output aids decision makers when allocating and assigning resources to areas of concern. The model not only identifies high-risk areas, based on the algorithm described in Appendix B, but also the repair rate, measured per 1000 feet of pipe. The repair rate is adopted from the American Lifeline Alliances (ALA) damage model for water pipes and takes into consideration damage equations as a function of ground shaking (peak ground velocity measured in inches per second) and ground failure (permanent ground displacement or PGD, measured in inches) [3]. The report itself is generated within 15 minutes of an event, without review from the Geosciences Department. It is immediately distributed to Gas Emergency Preparedness and Readiness (EP&R) and Integrity Management personnel via company email. Within 90 minutes, a reviewed DASH report is distributed for any earthquake rated as M5.0 or greater. For events measuring less than M5.0 (but greater than M3.0) the report is only made available through the DASH homepage on the PG&E intranet [29]. Throughout the life of an event, the DASH output can be revised and re-sent if significant changes are found. For reference, a 19th version of the DASH report was sent out approximately 48 hours after the start of the 2014 American Canyon earthquake.

The DASH model is an important tool for Integrity Management personnel, but it is not the only input analyzed when considering potential areas of pipe failure. Following an event, DIMP also reviews additional data sources to better predict which plats need to be surveyed. According to the Asset Knowledge and Integrity Management (AK&IM) "Playbook," other important sources include police, fire department, and customer phone calls through dispatch and customer care and billing (CC&B), building damage through online sources and local city governments, and historic survey schedules and leaks from DIMP internal resources. There is no clear method for gathering these data or protocol for sharing the above information. Instead, the above items are listed in appendix I of the "AK&IM Playbook" and are meant to help guide DIMP decision makers and GEC personnel as a supplementary checklist.

Since PG&E has extensive knowledge and experience understanding the geological features of their service territory, we chose to build on the framework they have developed by incorporating information that is readily available and already recommended for consideration. The system our model augments currently involves a manual, human assessment of the data and can be taxing for even the most experienced decision maker. Using DASH as the core feature, while including additional variables into an inclusive model, we aimed to improve the predictive performance.

An additional benefit to building on the existing framework, is that the model output was already providing information on a per plat basis, and the front-end interface was familiar for emergency response employees. Emergency management relies on significant training and familiarity with the tools available. Large-scale changes without adequate time and resources to retrain personnel could result in degraded

response efforts. With that in mind, the idea of gradually introducing a new system was an important characteristic of the newly proposed damage prediction model.

3.3 Model Assumptions and Limitations

Analysis of the dataset and discussions with PG&E personnel drove us to identify several limiting factors and make a series of assumptions. The most important limitation involves the historical data used to train the model. We were limited on the information available to the 2014 American Canyon earthquake. Information from the 1989 Loma Prieta earthquake was restricted to macro level after-action reports and personal accounts. The information gathered from the 1989 event influenced predictor variable choices but could not be relied upon for plat level accuracy when building historical datasets for model training. Furthermore, certain areas damaged in the 1989 earthquake resulted in line replacement instead of repair. Individual leaks were not catalogued on pipes designated for replacement because it was seen as a wasted effort if the line was to be removed. As a result, extensive data regarding leak reports are missing from that event.

Secondly, using information from a single earthquake assumes future earthquakes will act in a similar manner. This is an important assumption to note because the 2014 American Canyon earthquake registered as a M6.0, which is considered relatively mild in comparison to future earthquakes projected for the region. As a comparison, the 1906 San Francisco earthquake measured M7.9, which released approximately 700 times the energy as the 2014 American Canyon earthquake (as denoted by the expression $10^{(1.5*(7.9-6.0))}$ used for evaluating magnitude relationships). Developing a model based on a single event of this magnitude may reduce robustness.

Thirdly, historical USGS ShakeMap data used for model development came through the DASH report and was updated multiple times throughout the 2014 American Canyon earthquake. Generally, as more sensor data becomes available and field inspections identify various points of ground disruption, the estimates are revised to reflect the most up to date information. The 2014 American Canyon earthquake created problems for geologist because it occurred along an unknown fault. The model assumes future earthquakes will occur along more studied faults, resulting in more accurate results at an earlier time. To account for this our model uses DASH data from version 19, which came out 48 hours after the start of the event. With this assumption we maintain that PG&E will have more accurate USGS data immediately after an earthquake occurs during future events.

Lastly, the model assumes that internet will be available during the earthquake. Multiple datasets and field updates are required to generate the necessary inputs, and without server access and an ability to distribute data the model cannot be used. The current DASH model is more advantageous in this situation. DASH inputs can be simulated through USGS estimates of likely scenarios, of which multiple simulations have been run and damage reports produced. In the event of limited technological availability, PG&E personnel can refer to a scenario that most closely reflects the current situation, and reference hard copy print-outs while guiding initial emergency responses. The dynamic model description later in this chapter and future recommendations appearing in chapter 5 will make it apparent why our proposed model is not afforded the same simulation capability.

3.4 Pre-processing Data

Prior to model development we had to further process the dataset to make it usable for our learning algorithms. Clustering methods that use Euclidean distance measures and others that learn feature weights based on an optimized gradient descent algorithm (logistic regression and support vector

machine) are sensitive to feature scales. Using the StandardScalar function available through the Skelarn package in Python, the features were rescaled to reflect a mean of 0 and a standard deviation of 1 [30].

Missing values also required attention. As with feature scaling, tree-based algorithms are insensitive to such issues, but other potential algorithms need complete datasets in order to function. Simply eliminating plats with incomplete data was not considered based on the relatively small dataset currently available. Standard imputation methods often involve replacing missing numbers with the mean or mode of the attribute values, or estimating a distribution from the current data, and assigning and replacing the missing value accordingly [31]. We chose to use a k -nearest neighbor impute approach because the data is organized geographically. For example, attributes present in row 1 are assumed to be similar to attributes presented in row 2 based on their proximity. The plats in which they represent are typically adjacent to one another. By looking at the data from three adjacent plats (k=3, using a Euclidean distance metric) we replaced the missing value with the average data from the three nearest rows.

3.5 Classification and Regression Trade-offs

Understanding we had historic leak data available from the most recent earthquake, we decided to construct a model using supervised learning techniques. From this point we could examine classifying the number of leaks into group memberships, or treating the number of leaks as a continuous, target variable. The latter would provide a highly-specific output for use in the newly developed resource optimization model and could be accomplished through regression analysis.

A graphical analysis of the relationships between various features and the data distribution using a scatterplot matrix was initially used in an exploratory effort to identify possible relationships. With no obvious results, we implemented a random forest regression, as it is less sensitive to outliers and requires minimal tuning. In the random forest, a series of individual decision tress are constructed and the final output is the averaged result of the individual predictions. With this algorithm, not all features are considered while determining the best splitting point, but instead, a randomly selected subset of features is evaluated over multiple iterations. The increase in bias is offset by the reduced model variance (through averaging), thus yielding better results in comparison to individual trees. Branches are then split by minimizing the mean squared error (MSE). This value is the average of the sum of squared errors (SSE) between the predicted outcomes and the actual number of leaks [32, 33] shown in Equation (1).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \quad (1)$$

Where, N is the number of points,
$\hat{y}_i$ is the predicted value, and
y_i is the actual value of point i

The coefficient of determination (R-squared) for the initial trials showed poor results. The mean value representing fit to the regression line was 0.465 with a standard deviation of 0.024. This was further characterized by extreme overfitting, as the mean training set value was 0.907 with a standard deviation of 0.004. While a continuous value could be beneficial to the organization, company representatives encouraged developing a model with grouped outputs, consistent with DASH. Knowing the exact number of leaks would not drastically alter response efforts but relaying the data in a tiered manner could impact resource allocation.

Grouping results into bins also serves as a positive reminder of the level of uncertainty present in the predictions. A 2015 review of the distribution system damage with the predicted DASH results from the 2014 American Canyon earthquake recommended exercising caution when interpreting model results. In the executive summary it states that precise repair locations cannot be predicted and estimates should be "used to guide regional repair planning" [15]. While the aim of the proposed model is to improve upon the current system, the task is challenging based on the inherent uncertainty and limited training data.

For our model structure, the predicted number of repairs was initially classified into five separate categories: 0, 1, 2-3, 4-5, and 6-10. In order to validate our best assumptions for bin sizes we held interviews with gas control personnel and subject matter experts. Following these discussions bins were reduced to three distinct groups, as displayed in Table 3.1.

Table 3.1: Bin numbers and corresponding number of leaks

NUMBER OF LEAKS	BIN ASSIGNMENT
0	0
1-5	1
6+	2

The model assigns each plat with a value of 0, 1 or 2, corresponding to 0 leaks, 1-5 leaks, or 6 or greater leaks respectively. Through interviews we determined plats experiencing greater than 5 leaks would be candidates for a shut-in, which is consistent with our upper-level bin choice. A shut-in allows PG&E to turn-off gas services to entire Emergency Shut-in Zones (ESZ). These zones can be as small as three plats but can affect 10,000 customers at any given time. Therefore, the decision to use them often requires catastrophic damage.

Once bin values were adjusted and assigned, the decision was made to use a classification algorithm to predict the damage (in number of leaks) on a per plat basis. The end result of the model was designed to display the plats affected by the earthquake, and a value (0, 1, or 2) corresponding with the predicted level of damage.

3.6 Performance Metric for Algorithm Comparison

In order to assess the performance of individual algorithms, we had to decide on an appropriate scoring metric. A preliminary analysis of the data showed a large imbalance between the classes. The approximate breakdown is 88% class-0 membership, 8% class-1 membership and 4% class-2 membership. A lazy algorithm could achieve an accuracy score, defined as the number of true predictions divided by the total number of predictions, of 88% by always predicting class-0 (assuming a similar distribution in the data generated from the next earthquake). As is typical with datasets exhibiting heavy imbalance, accuracy was not considered as a representative performance metric.

Alternatively, precision and recall values were examined to assess model performance. Recall (REC) is defined as the number of true positives divided by the total number of positive samples. Precision (PRE) is defined as the number of true positives (TP) divided by the sum of true positives and false positives (FP) [34]. Since our model was built on a multi-class classification problem, traditional terminology may be inaccurate or misleading. In our example, a false prediction is instead noted as an "error."

The confusion matrix below (Fig. 4) provides an illustrative example of how model results are displayed. The term TP represents a "true positive," or a correct prediction. In this context, a leak is neither negative

or positive, but instead, the prediction is positive (correct) or it is an error (incorrect). The subscript refers to the class. Consequently, the term E represents an "error," and the subscript refers to its row/column position in the confusion matrix. For example, $E_{0,1}$ refers to an error, where the true value is a class-0, but the predicted value is class-1. Using Fig. 4 as an example, the PRE of class-1 is defined as:

$$\frac{TP_1}{(TP_1 + E_{0,1} + E_{2,1})} \tag{2}$$

and the REC of class-2 is defined as:

$$\frac{TP_1}{(TP_2 + E_{2,1} + E_{2,0})} \tag{3}$$

$$\begin{array}{cc} & \textbf{\textit{Predicted Label}} \\ & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \textbf{\textit{True Label}} \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{bmatrix} TP_0 & E_{0,1} & E_{0,2} \\ E_{1,0} & TP_1 & E_{1,2} \\ E_{2,0} & E_{2,1} & TP_2 \end{bmatrix} \end{array}$$

Figure 4: Confusion Matrix for 3-class classification algorithm

Often times the precision and recall scores are combined through their harmonic mean as a better assessor of performance. This is known as the F1 score and is defined as twice the product of (2) and (3) divided by the sum of (2) and (3) [34].

However, this particular case required an additional level of detail not satisfied through the use of traditional recall, precision, and F1 scoring metrics. In the example described above, the class-2 recall score does not differentiate between $E_{2,0}$ and $E_{2,1}$. In practice those errors do not result in the same consequences. For example, predicting a plat had between 1-5 leaks (class-1) would warrant a response from emergency personnel. Upon verification that the scene is more damaged than initially predicted, additional assets can be requested. Conversely, if a plat is predicted to have 0 leaks (class-0), when more than 6 leaks are present (class-2), gas will be released over a greater period of time as no assets would be directed to respond. The longer gas is released, the greater the probability for it to develop into an inhalation, fire or explosive hazard. We aimed to minimize the time of gas blowing, and therefore, did not assume all false predictions carried the same weight.

To account for this factor, individual results from the confusion matrix were applied against a penalty matrix. Consequences of a false prediction were assigned weights on a ten-point scale, penalizing errors that fail to predict a leak more than predictions that incorrectly assign a leak(s) to a plat when none are actually present. Penalty severity is then determined based on the consequences of the error. Misses between adjacent classes are less harmful then failing to distinguish between a class-0 and a class-2. In the model, the results of the confusion matrix were unraveled and presented as a [1x9] matrix. The dot product was then taken between the unraveled confusion matrix and a [9x1] penalty matrix, resulting in a final score. Algorithms with lower scores are considered better performers. An example of this process is shown in Fig. 5.

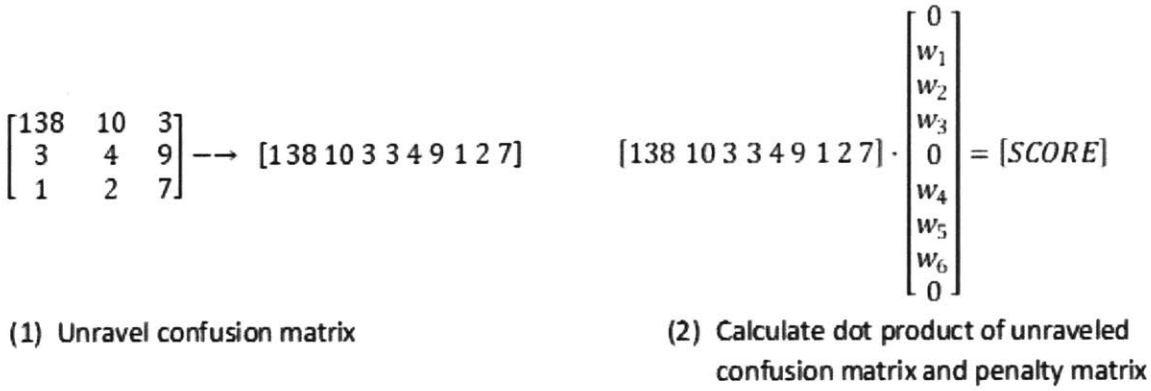


Figure 5: Confusion matrix applied against penalty matrix to evaluate algorithm performance

The penalty matrix for this model assigns a 0 to all true predictions. The model is never penalized for a correct prediction. w_1, w_2, and w_4, are assigned a value of 1 because they over predict the damage. From a safety standpoint this is not a bad result. An argument can be made that resources are being used inefficiently, but the plats they are assigned will require inspection at some point regardless. Therefore, they have not been penalized harshly. w_3 is assigned a value of 5 because it fails to predict leaks categorized as class-1 (1-5 leaks), and w_5 is assigned a value of 10 because it fails to predict leaks categorized as class-2 (6 or more leaks), with the latter being a more egregious error. Finally, w_6 is assigned a value of 1 because it still predicts damage, but to a lesser extent. Assets will still respond to the plat and upon further inspection can begin mitigating the problem or requesting further assistance. The weights were chosen based on input from gas operations subject matter experts. The penalty matrix described is shown below:

$$[0\ 1\ 1\ 5\ 0\ 1\ 10\ 1\ 0]^T$$

This scoring metric will vary greatly depending on the number of samples tested. Therefore, it must be noted that It is only used as a comparison tool in evaluating the various learning algorithms.

For model evaluation the data was split into training and testing sets using 10-fold cross validation. This involves randomly splitting the data into 10 sets without resampling, using nine folds to train and saving the final fold for testing. This process is repeated until 10 separate models have been produced. The performance is then assessed as the average result of the 10 models. This method is advantageous to holding out a testing set because it becomes less sensitive to data partitioning. To further improve the evaluation process, we also utilized a stratified method to account for label imbalance. This ensured that each fold preserved the class proportions represented in the entire dataset.

3.7 Static Model

Initial model development was based on a static framework. We made the assumption that all of the data required to make predictions regarding plat damage was immediately available at the onset of the event. By treating the problem in this manner, we could work backwards in time in order to identify critical decision points. Starting with a baseline performance (prediction power with all the data available), we

could selectively remove points of information based on their actual availability in time to determine model impact. This process is described in detail in section 3.8.

Predictor variables were separated into distinct categories as described in Table 3.2. This organizational method was based on the different datasets from which the information was available and allowed us to select and discard particular features, while still maintaining categorical representation. Appendix E further provides detailed explanations for each feature that was explored and its source.

Table 3.2: Categories and associated features used in developing the predictive model

CATEGORY	FEATURES				
DASH	EP	PGA	LS	LQ	Fault Value
HISTORICAL SURVEYS	Grade	No. leaks	Time Since Last Survey	-	-
CENSUS	Population	No. Buildings	Building Type	-	-
INFRASTRUCTURE COMPONENTS	Length	Material	No. Service Taps	-	-
BUILDING DAMAGE	HAZUS	Reported	-	-	-
COMMUNITY DECIMAL INTENSITY (CDI)	Value	No. Responses	-	-	-
CALLS	Customer	PD/FD/EMT	-	-	-

Where,
Length = Pipe length in feet
Material = Pipe material
No. Service Taps = Number of above ground service taps
Grade = Leak severity from most recent survey based on [8]
Time Since Last Survey = Number of days since the plat was last surveyed. PG&E surveys plats on a scheduled basis for leaks in order to ensure they continuously provide safe and reliable services. Plat surveys range from 1 to 3 years.
No. Building = Number of buildings located within the plat
Building Type = Building inventory has been broken into 33 distinct types per the HAZUS-MH building inventory used by FEMA for predicting building damage. Major categories such as residential, community, industrial, agriculture, government and education buildings are further divided to reflect their size and construction nuances.
EP = Earthquake Prioritization Value (See Appendix B)
PGA = Peak ground acceleration
LS = Landslide susceptibility (See Appendix B)
LQ = Liquefaction Susceptibility (See Appendix B)
Fault Value = If the plat falls along fault line and earthquake magnitude is greater than 6.0, a 1 is assigned, else 0.
HAZUS = Number of buildings damaged as predicted through HAZUS-MH software
Reported = Number of buildings damaged as reported from actual observers
Value = CDI value
No. Responses = Number of individuals responding through DYFI program
Customer = Number of received phone calls by PG&E for emergency or odor concerns from customers
PD/FD/EMT = Number of calls from Police Department (PD), Fire Department (FD) Emergency Medical Technicians (EMT) calling to report gas related issues

Each category is composed of multiple features that could be used individually, in total, or in combination in the predictive model. As the hypothesis states, they have been chosen based on previous research, after-action reports, and company interviews. For example, under the "Building Damage" category we can receive information through the FEMA developed HAZUS software and/or reports from city and state driven damage surveys. For our model we examined the Earthquake Engineering Research Institutes (EERI) post-earthquake reconnaissance survey for the 2014 American Canyon earthquake to represent "reported" damage. This information was used to identify buildings that were damaged, the extent, and where they were located. In actuality, these data would not be available for several days after the event, but we decided to examine its importance and impact on the model's predictive power. If the predictive software (HAZUS-MH) accurately reflected the actual damage, then these predictions could be used as a feature within the damage model pipeline. This is based on the hypothesis that building damage increases the probability of above-ground pipe damage. Ultimately, it was not included in the final model, and an explanation is presented in chapter 5. Final features chosen for model development are presented in Appendix D.

Once the basic structure of the dataset and model was complete we could begin looking at simulating real-world events. We examined how data would become accessible, at what point in time, and how it would affect model results. This was accomplished through the dynamic model.

3.8 Dynamic Model

The categories chosen for the prediction model can be separated into three distinct periods based on their availability: prior to the event, at the onset of the event, and post event.

Prior: This includes knowledge of the infrastructure, previous leaks through historical surveys, and census data regarding population and building types. Without knowledge of the earthquake characteristics, such as intensity and location, very little can be inferred from these categories alone. Basic assumptions remain valid, such as more service taps and more pipes increase the chance of multiple leaks in a single area. Additionally, plats that have historically exhibited an above-average number of leaks (whether it is based on soil corrosiveness, pipeline material, sheer number of assets, etc.) or those that have not been surveyed in several years are more likely to exhibit leaks post-earthquake. Independent of whether they were caused by shaking or ground disturbances, these leaks will result in PG&E response efforts and resource expenditures.

Onset: After an earthquake, PG&E utilizes data from USGS ShakeMaps, which provide "near real-time maps of ground motion and shaking intensity following significant earthquakes" [35]. These data provide the input for internal company damage prediction algorithms, including peak ground acceleration, velocity, and disturbance. These same values are also available as input for building damage models available through the HAZUS-MH software. Current PG&E damage modeling efforts (excluding the use of HAZUS-MH) would typically conclude at this point. Once all of the USGS data has been received as input for DASH, normally 60-90 minutes after the event [29], additional predictive efforts would seize. Only if significant changes were reported would these models be amended and redistributed throughout the organization (as seen through DASH version 19).

Post: Throughout the emergency response, PG&E receives phone calls from both customers and first responders regarding known and perceived gas related incidents. Calls range from fire department personnel requesting to secure gas services to customers misinterpreting garbage odors as hazardous gas leaks. Additionally, USGS provides an online platform for individuals to report earthquakes through an online survey. Results from the "Did You Feel It" (DYFI) questionnaire are inputted into an algorithm to

estimate event intensity. During the 2014 American Canyon earthquake, this program received over 40,000 responses. As time progresses, more data points are produced and can be used to help localize areas experiencing significant damage.

The dynamic model takes into consideration all three periods over an iterative process. Streaming data from the 2014 American Canyon earthquake has been segregated into 2, 4, 8, 12 and 24-hour periods. This includes calls received and USGS questionnaire responses. New developments force the algorithm to update the weights assigned to the independent variables. As a result, the final product can be described as series of individual, time-dependent models, with each taking advantage of the most up-to-date reports. In this case, small datasets also provide a useful advantage. PG&E's entire service territory is divided into approximately 22,000 separate plats. During an earthquake, only as small fraction of these will be affected. Prior to any dimensionality reduction, the resulting datasets still remain modest in size and can be executed without considering time delays or limits in computational power. This not only ensures multiple iterations can be run over the life of the event, but if desired, the model can be re-trained in real-time using information from plats with confirmed leaks.

A more detailed analysis is described in the results section, but early trials showed incorporating updated data from field personnel and customers improved model performance. Interestingly enough, the greatest impact is seen within the first two-hours of the event. This enables PG&E to use this information to develop an action plan and request mutual aid in the same amount of time it takes to activate the emergency control centers.

3.9 Algorithm Selection

There are a multitude of learning algorithms available to choose from and each one maintains its own strengths and weaknesses. Having created the necessary datasets segregated over specified time periods after the event, we examined the predictive capabilities of logistic regression, support vector machine (SVM), k-nearest neighbor (KNN), and random forest models.

Logistic Regression: This classification model predicts the probabilities of class labels. To accomplish this, the model establishes a linear relationship between the predictor variables and the inverse of the log-odds ratio. The odds ratio is defined as the probability (p) of an event divided by the quantity one minus the probability of an event. This can be written as

$$\frac{p}{(1-p)} \tag{4}$$

By taking the logarithm of this ratio, input values are unrestricted and can span the range of real numbers. However, since the model needs to determine the probability, the inverse of the log-odds ratio is required. This can be written as

$$\frac{1}{(1+e^{-z})} \tag{5}$$

$$\text{where } z = \sum_{i=1}^{1} x_i . w_i \tag{6}$$

The output of the sigmoid function, or the probability of class inclusion given features x parameterized by w, then becomes the cost function of the model.

The weights are then determined using a gradient ascent (or descent) optimization algorithm by maximizing (or minimizing) the log-likelihood (cost) function described above. The log of the equation makes determining the partial derivative with respect to each weight much easier. This allows us to identify the gradient and update the weights by moving away until reaching the global maximum (or minimum).

Support Vector Machine (SVM): SVMs are best described visually using a two-dimensional feature space, with each feature value represented as a specific coordinate. Fig.6 shows two classes separated by a decision boundary or hyperplane. In this example, you can visualize how multiple hyperplanes could exist that would still effectively separate the two classes. In order to determine the optimal boundary, which can be defined as

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{7}$$

Where, $\mathbf{w}^T$ is the weight vector
and $\mathbf{x}$ is the input vector

the algorithm maximizes the distance between the boundary and the support vectors. The vectors are the individual coordinates that lie on the hyperplanes defining our margin. In this case, the colored-in items represent support vectors used to determine the maximum margin.

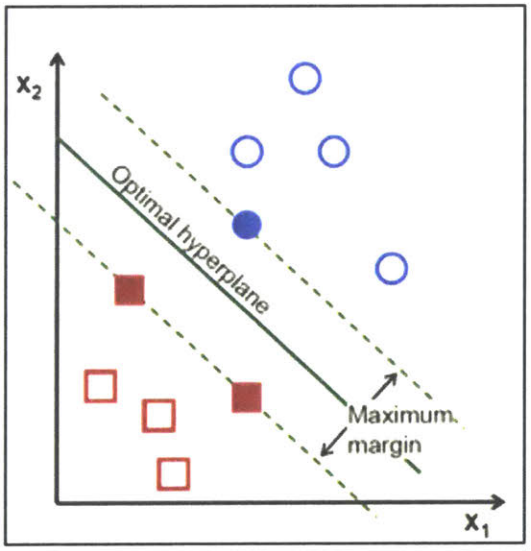


Image taken from OpenCV and be accessed at http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

Figure 6: Graphic representation of support vector machine

The margin is defined as $\frac{2}{\|\mathbf{w}\|}$, which can be shown through Fig. 6 as well. The dotted lines represent the "positive" and "negative" hyperplanes and are defined as:

$$\mathbf{w}^T\mathbf{x} + b = 1 \text{ and} \tag{8}$$
$$\mathbf{w}^T\mathbf{x} + b = -1 \tag{9}$$

Samples are categorized as -1 and 1 instead of 1 and 0. By subtracting the two equations and normalizing by the length of vector $\mathbf{w}$, written as

$$\sqrt{\sum_{j=1}^{m} \mathbf{w}_j^2} \tag{10}$$

we have defined the margin in which we look to maximize.

Non-linearly separable problems can also be solved with SVMs using a kernel function. This function transforms the data to a higher-dimensional feature space. Once transformed by a mapping function, the data becomes linearly separable.

Both the logistic regression and SVM classifier use a "one versus rest" technique in order perform on multi-class problems. In this way, individual classes are treated as the positive sample, and all other classes are grouped together to represent the negative class. This process is repeated for all individual classes, where a confidence score is produced and then compared against other iterations prior to determining the sample class.

K-Nearest Neighbor (KNN): The KNN algorithm is a non-parametric model that requires the use of the training set for each new prediction. Unlike the logistic regression model and linear SVM mentioned above, this algorithm does not construct an internal set of rules. Instead, the training data is stored and used to classify new instances. This memory-based approach can be computationally intensive for large datasets but was appropriate for this work.

Each new instance is compared against a number (k) of neighbors in n-dimensional space, depending on the number of predictor variables chosen. The theory is that instances surrounded by a majority of one class will belong to the same class with a high degree of probability. There are many different metrics available for determining the closest set of neighboring data points, but the most common method is to use the Euclidean distance. This method requires the dataset to be normalized in order to ensure specific predictor variables are not disproportionally influencing the classification.

Random Forest: The final algorithm explored in the project is a series of decision tree classifiers that have been built from randomly selected samples of the initial training set. These new training sets are assembled and compared against a randomly selected number of predictor features, resulting in a series of trees with unique predictions. The results are then combined with the class receiving the majority of votes chosen for the ultimate prediction.

Decision trees themselves can be thought of as a way to partition the dataset, such that each decision point maximizes the purity of the subsequent set. For example, if node "A" is comprised of 10 samples belonging to class-1 and 10 samples belonging class-2, it would be very difficult to make a prediction. However, if node "A" is split based on an additional criterion resulting in sub-node "B" with only class-1 membership and sub-node "C" with only class 2 membership respectively, the problem is much easier as the new nodes are considered pure. With real-world datasets, it may be unreasonable to expect pure nodes from both a computational and time perspective, and as a result, algorithm parameters can be set to establish a maximum number of nodes along with a purity threshold.

The random forest can be implemented very easily because it does not require any data pre-processing. It also takes advantage of a series of predictions, that by themselves may be flawed, but when aggregated can overcome their individual errors.

Preliminary test results using the four algorithms indicated over-fitting. This was expected for the KNN and random forest based on the nature of the algorithms but was also evident while comparing training and test set performance with the logistic regression model and SVM. Outside of cross-validation techniques, it was recognized that potentially irrelevant features could also be eliminated to help reduce over-fitting. Earthquakes in the San Francisco region resulting in major damage to the natural gas distribution system are very limited and data from these events are often missing or non-existent. The end result is a relatively small set of data points with a large number of possible predictive features. Reducing the ratio of features to observations was important for generalizing the results.

At this point it is important to comment on the project holistically and summarize the overall process. As previously noted, quantitative and qualitative information was collected from engineering damage assessments and employee interviews to media reports and customer phone calls to best determine predictive features. The approach to this initial phase was to collect as much information as possible while on site, allowing for model refinement to occur at a later date, independent of location. The data was then compiled into a useable set with the understanding that specific features could be removed throughout the process with various selection techniques. Alternatively, features could have been added sequentially, but the former allowed for the examination of all the features together. This process was seen as a better alternative to guessing which additional predictors would have the largest, positive effect.

Many methods exist, and for this work both sequential backward selection (SBS) and a random forest were used for identifying relevant features. SBS sequentially removes a predictor variable and creates a new classifier with the subset of features. This process continues and each classifier is evaluated against a pre-determined metric, such as predictive performance or error rate [36]. The random forest classifier helps reduce the number of features by looking at the decrease in impurity across all individual trees for a specific feature. This means that specific nodes, once split, are better at separating data points, and are therefore more effective at predicting the class of new observations. Once the impurity decrease values are summed the results can be ordered and normalized to indicate the most important features.

By using a Random Forest, we were able to examine the initially chosen predictor variables and identify the most useful ones in terms of predictive performance. A normalized graph displaying feature importance is shown in Fig 7. It is important to note, however, that the results were based on a single earthquake of modest magnitude. In an effort to make the tool as robust as possible, not all of the features were eliminated. This opens the possibility for new data to be effortlessly incorporated into the model during the next major event. A clear example involves the location of gas lines in relation to major fault lines. The American Canyon earthquake in 2014 occurred along a previously unknown fault, which is atypical and unlikely for future events.

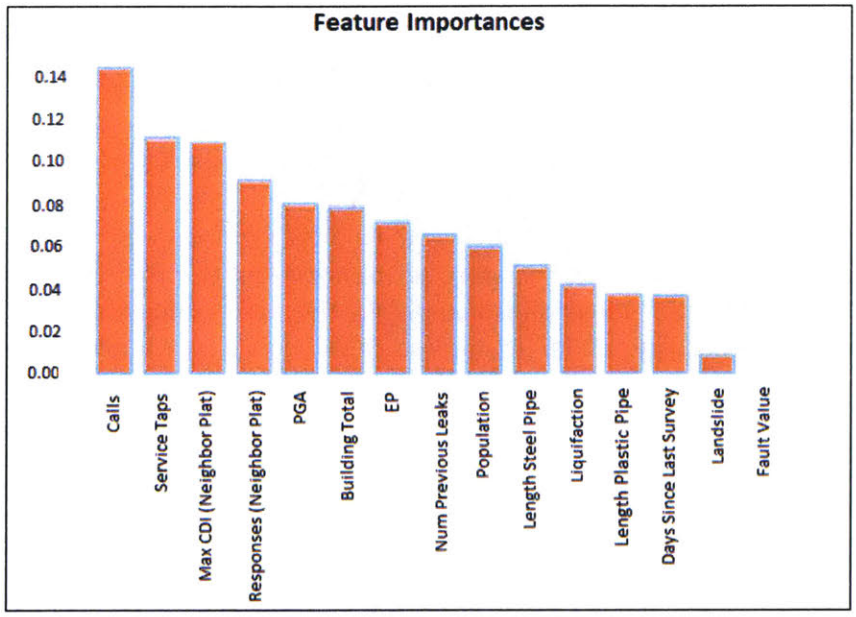


Figure 7: Example feature importances based on random forest algorithm

Following dimensionality reduction and efforts to maintain model flexibility, the dataset was broken into four categories (CAT I, CAT II, CAT III, and CAT IV). Appendix D shows the predictor variables included for each subset. The categories progress from I to IV in terms of implementation difficulty, which was based on the accessibility of the data (if it existed internally or was dependent on outside resources), pre-processing and formatting requirements, and historical quality. For example, CAT I could be used immediately based on PG&E's current system, whereas CAT IV would require the creation and maintenance of several new datasets, which could prolong implementation efforts.

Final model tuning was done through a grid search method. The moderate size of the dataset allowed for an exhaustive search, meaning all possible hyperparameter combinations (based on a predetermined set of possible values) were evaluated against the custom performance metric described in section 3.6. Hyperparameter examples include regularization values (logistic regression and SVM), number of neighbors and distance metric (KNN), and tree depth, number, and information gain function type (random forest). Using this method in conjunction with cross validation techniques yielded the optimized set of parameters for each model.

As mentioned in section 3.8, part of the overall damage prediction improvement project involved using real-time data to adjust initial predictions. This requires feature weights to change throughout the event, as specific information becomes more accurate and readily available. Therefore, the model can be interpreted as a number of unique, individual algorithms depending on the time into the event. In our case, we utilized four distinct datasets, that were trained and validated over five time-periods. Each algorithm was executed at 2, 4, 8, 12 and 24 hours after the start of the earthquake., resulting in 20 distinct models. These periods were chosen for several reasons, including internal company response metrics, recognizing the earliest information is the most volatile and requires the most attention, as well as the need to establish realistic time periods between running the model and using the data to direct resource allocation.

3.10 Results

The four learning algorithms were run against historical data from the 2014 American Canyon earthquake, using the custom designed scoring metric in order to evaluate performance. Trials were run 500 times for each CAT I-IV dataset at time intervals of 2, 4, 8, 12, and 24 hours. It should be reiterated that the scoring metric is dependent on the size of the data set and class population. For example, based on the penalty matrix, a dataset with twice the number of observations could perform better, but still incur a greater overall penalty and appear less valuable. Therefore, it was important to ensure that class representation was balanced throughout the training and testing phases. This ensured the score was indicative of the algorithm and not an imbalanced set. With this in mind, final scores were compared such that the lowest value equated to the highest performing model.

The KNN and random forest algorithms performed similarly and each consistently erred when predicting plats with one to five leaks (class-1). Overall, the algorithms heavily favored predicting zero leaks had occurred. Based on an emergency response scenario, it was important for the model to err on the side of caution. These prediction failures resulted in scores approximately 30 – 60% greater (worse performer) compared to the logistic regression and SVM model scores.

The latter two models performed similarly to one another and were chosen for further exploration. Ultimately the logistic regression model was chosen based on overall performance and the immediate interpretation of results. Each outcome was assigned a probability, which could be relayed to the decision maker. Plats could then be sorted based on the interpreted confidence level, and leak survey assignments adjusted if outside information not present in the model became available.

The logistic regression model was used with $L2$ regularization and inverse strength of 0.01 to 0.1. Depending on the time into the event and the dataset being used, different hyperparameters were chosen through a grid search optimization function. Treating the overall model as series of algorithms executed throughout different points of the earthquake response allowed for adjustments to be made in line with the arrival of new information (characterized by the dynamic model). The average score of the 20 individual logistic regression models (based on category of predictor variables and time interval) was 37.93 with a standard deviation of 0.65. As indicated scores were relatively close, with a difference of 2.46 between the best and worst performing algorithm.

Section 3.9 discussed the partitioning of data between four categories, which was done to provide options based on the level of difficulty and effort required to implement within the organization. At that point it was hypothesized datasets with additional information would perform better, but this was not the case. While CAT II features at 24 hours into the event provided the best predictive model, CAT I features were overall the most useful. Adding features during later time intervals did, however, lower the standard deviation for the scores produced throughout the 500 trials.

Overall the algorithms using datasets from later time intervals did prove to be more effective in mean performance. For example, algorithms trained with historical data that was available at the 24-hour mark performed better as a whole, but less significantly than anticipated. Incorporating real-time data is most effective in the first 2 hours following an event, at which point the benefit begins to taper. Full results are shown in Table 3.3 below:

Table 3.3: Results of the logistic regression model at various time steps using L2 regularization

Time	CAT I		CAT II		CAT III		CAT IV	
	Avg. Score	Std.	Avg. Score	Std.	Avg. Score	Std.	Avg. Score	Std.
2 Hrs	37.592	8.761	39.256	9.677	38.622	9.769	39.256	9.007
4 Hrs	37.724	8.426	37.938	8.964	37.106	8.470	38.738	9.391
8 Hrs	37.621	8.275	38.148	9.312	37.818	6.439	37.920	6.282
12 Hrs	37.232	8.744	37.820	9.982	37.468	6.656	38.476	6.157
24 Hrs	37.698	8.694	36.800	9.870	37.254	6.480	38.082	6.515

Establishing a method of comparison between the proposed model and the one currently in place is difficult because the two models do not share the same performance metrics. For example, the current system identifies a value known as an "earthquake prioritization" (EP) number. This number is derived from an internally developed algorithm that considers geological factors (peak ground acceleration, landslide and liquefaction susceptibility, and fault location) along with pipe material. The plats are then ranked from highest to lowest, with the higher value indicating a greater likelihood for damage. High-risk plats above a company determined EP threshold are then surveyed for leaks. Additionally, this system only takes into consideration main distribution lines, and does not consider branch lines or meter sets in the assessment. The proposed model considers all portions of the distribution system, while incorporating features of the current systems as a subset of predictor variables.

Based on this, measuring model impact can be done in two specific ways. The first method treats the newly developed model as if it were currently being utilized by PG&E. To simulate this scenario, the main input currently used by PG&E (EP value) was used in our proposed model as the sole predictor variable. This resulted in a mean score of 80.798 and standard deviation of 8.128, marking an approximately 74% degradation in performance when compared to our proposed model with additional predictor variables. The EP values used in the historical dataset were from the 19^{th} iteration of DASH, which occurred 48 hours after the onset of the event. This was done in a conservative effort to account for irregularities in the earthquake. The assumption was future earthquakes will occur along known fault lines, which will in turn lead to more representative EP values within the first 90 minutes following an event.

Further building on this method, we can also assess the benefits of the overall model development and addition of dynamic variables. By treating features used to calculate the currently used EP value as individual predictors instead, we can see how the model design improves upon current performance. For example, extracting pipe material and PGA and using them as unique parameters instead of features of the EP algorithm allows us to see immediate improvements. This technique yields a means score of 40.261, which is an improvement of approximately 67% from relying strictly on the EP value. Furthermore, when comparing the mean score of 40.261 with the average performance of the 20 models (37.592) we can assess the benefit of adding dynamic data. Here we see an improvement of 6.8% as such predictors as phone calls are included.

The second method of comparison shows a visual representation of the model output as it would relate to emergency response efforts. The value of the model is derived from the amount of time saved in identifying potentially hazardous leaks. Therefore, it is beneficial to examine the time periods and success rates for identifying plats with leaks. The implementation of the model and how it relates to PG&E's emergency response operations will be discussed in detail in chapter 4.

For reference, Fig. 8 shows how the service territory is covered between individual plats. The left most image depicts PG&E service plats as they relate to the Sonoma and North Bay Divisions. The image to the right represents the area of interest for this study with the satellite imagery removed for clarity.

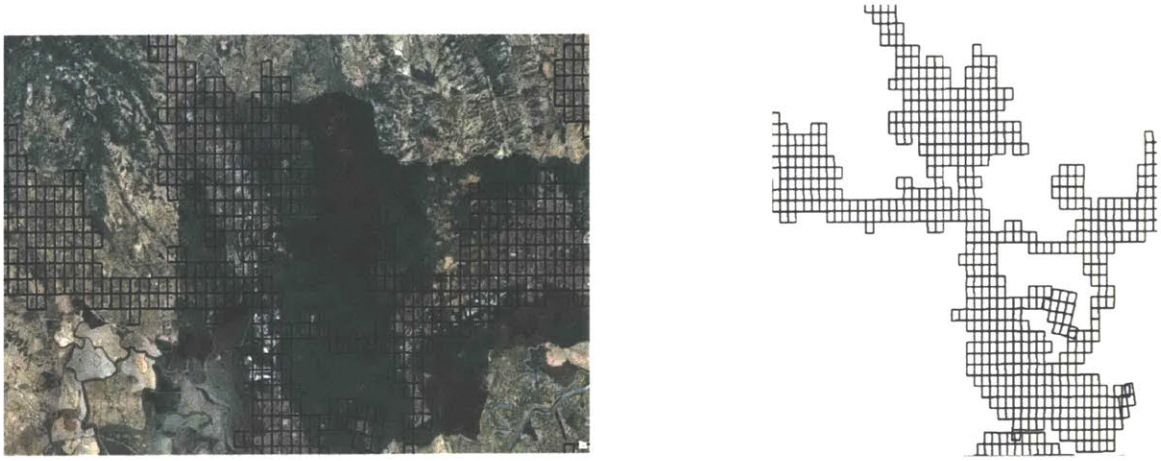


Figure 8: Plat map overlay with service area satellite imagery and plat map with background removed

For illustrative purposes, 129 plats (with a balanced class representation) were randomly selected for a test dataset. The predictor variables came from CAT I data collected two hours after the onset of the event. The set contained 114 true class-0 plats, 11 true class-1 plats, and four true class-2 plats. The breakdown is shown in the confusion matrix in Fig. 9 below.

		Predicted Label		
		0	1	2
True Label	0	102	11	1
	1	3	6	2
	2	0	1	3

Figure 9: Confusion Matrix of model results from selected trial

Fig. 10 visually displays these results as they apply to the location of specific plats. In this simulation 12 of the 15 plats were identified as having leaks by the proposed model. Three were correctly identified as having leaks but were predicted as the wrong class. 12 plats were erroneously classified as having leaks when they were true class-0s. Fig. 10 showcases the results using a color-coded map.

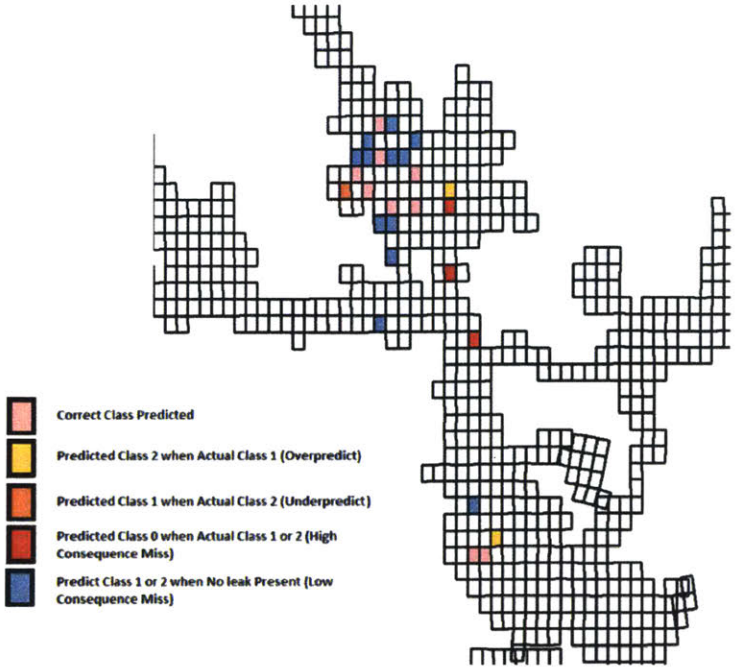


Figure 10: Color-coded plat map results for proposed model

In comparison, Fig. 11 shows the predicted results used by the DASH model and the resulting leak survey response during the 2014 American Canyon earthquake. In this real-world representation, seven of the 15 plats with leaks were correctly identified, eight plats were missed, and 11 plats were predicted to have leaks when none were actually present.

Figure 11: Color-coded plat map results from DASH model

This second method of comparison is useful because it translates the results from a confusion matrix and customized scoring metric into a real problem. By analyzing the simulated scenario using historical response efforts as a baseline, the value of the model is better understood. A summary of the results is shown in Table 3.4. While the proposed model predicts more plats than the past system, the number of plats surveyed for damage when none existed is extremely close in both cases. This precludes the argument that the proposed model identified more plats correctly based on the number of predictions and not on the methodology described throughout the paper. It is further important to note the proposed model's performance was evaluated based on information two hours after the onset of the event, whereas the actual model was evaluated over a course of 48 hours. PG&E's response efforts are not strictly based on DASH but take into account many other pieces of information. By examining the data after two days of emergency operations we have a better sense of the effectiveness of the current system as a whole. In that regard, it should still be noted that the proposed model has the potential to save a significant amount of time in deciding which plats should be surveyed following an event.

Table 3.4: Summary of results between proposed model and current model

DESCRIPTION	PROPOSED MODEL	ACTUAL MODEL
NUMBER OF PLATS IDENTIFIED FOR SURVEY	24	18
PLATS INCORRECTLY CHOSEN FOR SURVEY (NO DAMAGE)	12	11
PLATS CORRECTLY IDENTIFIED (PREDICTED AND CONFIRMED LEAKS)	12*	7
PLATS MISSED (DAMAGED BUT NOT PREDICTED)	3	8
TIME AFTER EVENT ONSET (HOURS)	2	48**

* 9 plats were predicted correctly by class; one plat was predicted to be class-1 when it was class-2, and two plats were predicted to be class-2 when they were class-1. The total of 12 is listed to reflect similarities with the actual model, which does not differentiate between the number of leaks when listing EP values that prioritize survey efforts.

** 48 hours represents the time an actionable survey plan was established, corresponding with the release of DASH version 19. It does not represent when plats were surveyed

4 Implementation of Model Results

4.1 Forming a Tool for the Decision Maker

In order to implement the model within the organization it was important to identify existing methods and resources that could be built upon rather than replaced. When dealing with issues as sensitive as emergency response, major changes often require extensive periods of retraining and validation. By using current training and readily available information, a new model could be streamlined into the current emergency response procedures.

Internal company literature produced from the Asset Knowledge and Integrity Management teams was heavily used throughout model development. The literature provides checklists to aid the Integrity Management Programs in identifying potential leaks. These lists further contain the source, format, and responsible parties for multiple datasets internal and external to the organization. The initial model, influenced heavily from said checklists, required manually pulling and formatting data from the various sources. Automating the entire process was critical, as time and resources could be saved by consolidating the efforts of multiple engineers and geologists across the company.

Once the series of algorithms were chosen and tuned, the model was rewritten to consolidate data from the Geosciences, Leak Survey, and Gas Distribution Control Center and formatted from Geographic Information Systems (GIS), Customer Care and Billing (CC&B), and SAP datasets. The information was used to create an independent variable matrix using the plat reference numbers as indices. In this way, each plat within the range of potential earthquake damage would have the predictor data immediately available. An example of this matrix exported as an excel file can be seen in Appendix F.

Determining a way to communicate the data to the appropriate parties for action was considered following these steps. Continuing development as an augmentation instead of a drastic overhaul of procedure, the chosen method for distributing the predictions was through the currently used DASH system. The project did not receive permission to alter the DASH file at the time of this writing, and so the resulting predictions were stored separately and designed to mimic the current report issued to the leak survey teams. A screen shot of the dataset is provided in Fig. 12, showing the class probability chosen for specific plats. The color-coding relates to the probability of class membership, with green indicating most likely and red indicating least likely.

Out[53]:

	Plats	Class0	Class1	Class2	Max
0	2639-H08	0.469501	0.22718	0.30332	Class0
1	2639-I08	0.564764	0.20957	0.225666	Class0
2	2639-J08	0.580021	0.209281	0.210698	Class0
3	2640-H01	0.544185	0.211449	0.244366	Class0
4	2640-I01	0.573376	0.211761	0.214863	Class0
5	2640-J01	0.561382	0.220607	0.21801	Class0
6	2640-J02	0.54376	0.226926	0.229313	Class0
7	2708-I7	0.624345	0.203368	0.172287	Class0
8	2708-I8	0.582381	0.230635	0.186984	Class0
9	2708-J6	0.570355	0.212577	0.217068	Class0

Figure 12: Final model output screenshot

The "Plats" column in Fig. 12 is a placeholder for the "Map #" column on the DASH report. Each class shows the probability for each prediction rather than a final prediction (as seen in the "Max" column in Fig. 12). This design was used because the algorithm is a tool for the decision maker and not a directive. Based on both the tremendous uncertainty involved with earthquake damage and cultural nuances within the organization, presenting data in this manner was ultimately decided as the preferred method. Chapter 5 will discuss future iterations on the design in detail.

SVMs can be presented in a similar manner, assuming confidence scores are required by the decision maker. In this case, probabilities have not been assigned, but the distance from the hyperplane has been inserted as a proxy for the confidence of the prediction. Using a "one-versus-rest" approach for multiclass classification, the results of the decision function in the scikit-learn (v0.19) library for Python can be ordered and sorted. With this information, Leak Survey teams can inspect the plats most likely to be damaged while awaiting further information to arrive and provide more insight on the state of other plats.

Additionally, a web application was designed to allow decision makers to query specific plats and receive a prediction. This design was a simple add-on feature to take advantage of the database of predictions, while providing information to members unable to receive the DASH report. Fig. 13 shows a screenshot of the application. The predicted class is shown along with a specific score. The score in this case is based on the probability of the prediction. Alternative scaling and normalization methods can be employed based on user desires. For example, a probability of 0.333 can be reflected as a score of 0 to display the ineffectiveness of the prediction for a 3-class model.

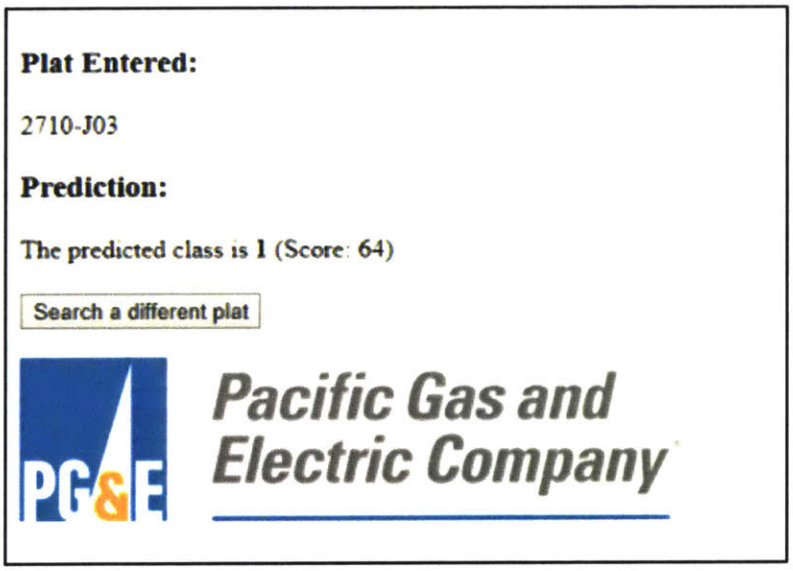


Figure 13: Web application screenshot

4.2 Resource Optimization

Beyond informing Leak Survey teams of plats with predicted leaks, results can also be funneled into resource optimization models currently being developed by Mohamed Kurdi and Bryan Hennessy [37]. Identifying how many assets need to be relocated from adjacent divisions to help with repairs is a difficult task. During an earthquake, local emergency center leaders often rely on past experiences when deciding the number of resources needed to restore service. Not only can this distributed approach be inefficient, it can lead to erroneous and costly requests for mutual aid from neighboring utility companies. Having estimates in the number of leaks over the entire service area can help leaders make more informed decisions. It has been proposed to use the model to average the number of leaks predicted in the area and then determine the required number of M&C and GSR crews required to address the damage.

4.3 Current Status

Taking the predictive model from theory to practice is currently on-going. DIMP has been identified as the most probable user of the data in order to help drive leak survey efforts. Outside of the Geosciences Department, they are the most familiar with the DASH reports and are able to integrate the results of the new model into current operating procedures. Secondary users include leadership at the EOC and OEC to help drive overall emergency response strategy. PG&E is continuously working to improve emergency response procedures and conducts, annual, multi-day company-wide exercises to prepare for future earthquakes. The predictive model was proposed at the August 2017 Exercise following observation of current operating procedures. The next step will be implementing the model within the 2018 training scenario in parallel with the DASH reports in order to compare the effectiveness. This means simulating multiple data sources and will require further analysis and research. Additional next steps are detailed in Section 5.

5 Conclusions and Future Work

5.1 Interpretation of Results

The newly proposed model shows promising results and we are cautiously optimistic about its future impact on PG&E's emergency response procedures. Recapping the efforts previously described, the model serves four major functions:

1) Expands current model to include branch lines and above-ground assets

2) Improves predictive power through the incorporation of additional features and real-time data

3) Eliminates manual procedures by automatically compiling data from multiple internal and external sources

4) Establishes framework for future model development

Historically, distribution main lines perform well during seismic events, which is why it was critical to include branch lines and meter sets in the model. With the removal of cast iron pipes in favor of steel and plastic, lateral movement induced from a seismic event (barring major ground disturbances) does not have a major effect on the structural integrity of the distribution system [38]. Additionally, the California Seismic Safety Commission describes the impact of building damage and pipe corrosion on gas safety during earthquakes, highlighting the need to incorporate additional predictors such as inspection dates and building density into the model. Thirdly, consolidating information into a single, searchable database greatly improves efficiency and leverages knowledge from across the organization.

The fourth feature, however, is arguably the most important model contribution. It has been noted multiple times throughout this document that the model was developed using information from a series of events spanning from 1906 to 2014, but quantitative data used for training and testing the algorithms came from a single event – The 2014 American Canyon earthquake. Records from the 1989 Loma Prieta earthquake (the most recent event above M6.0 to occur before the American Canyon earthquake in PG&E's service area) were not available for use. Therefore, the results of this model should be used cautiously with future events, which may also present peculiarities. However, ceasing damage prediction improvement efforts based on this fact was not considered because developing a framework to accept future observations could have a tremendous impact. The model is designed to continuously improve

based on new observations. Data from the next event can be immediately placed into the model, which will improve performance and lead to more powerful predictions in the future.

Further analyzing the model results we can also see an important financial benefit for the organization. The goal of the project was to ultimately reduce the time gas escapes from distribution assets following future earthquakes. This metric has been explored in chapter 3, which described temporal and spatial predictive improvements through the incorporation of additional predictor variables and real-time data. These enhancements to data driven operations improve safety and reduce the risk for critical infrastructure damage stemming from unnoticed or untended gas leaks. In order to quantify said improvements, historical incident costs in the state of California were examined over a 20-year period relating to infrastructure damage events that resulted in:

1) Fatality or injury requiring in-patient hospitalization
2) $50,000 or more in total costs, measured in 1984 dollars
3) Highly volatile liquid releases of 5 barrels or more or other liquid releases of 50 barrels or more
4) Liquid releases resulting in an unintentional fire or explosion

These criteria are determined by the Pipeline and Hazardous Materials Safety Administration (PHMSA), which establishes national policy, sets and enforces standards, educates, and conducts research to prevent incidents relating to the transportation of energy [39].

Following data compilation and appropriate geographic and asset filtering, we established a per incident cost of $284,423, given the event met the criteria above [40]. As the model was designed to identify leaks and not necessarily resulting incidents, a probability was assigned for each class indicating the likelihood a major incident would result. Based on historical notes and discussions class-1 bins were given a 0.5 probability of resulting in an incident, and class-2 bins were assigned a 0.75 probability. Using these values, each plat could be assigned a predicted cost based on the model classification.

Building on this methodology, plats were further assigned "vulnerability scores" (Vs) based on their characteristics. Highly-rural areas would be less likely to suffer damages as extreme as urban centers or areas with critical infrastructure. Three major categories were examined, to include:

1) Population
2) Gas Distribution Assets
3) High-Consequence Areas (HCAs)

The HCAs include schools, government buildings, hospitals, and emergency centers. Census data from 2012 was used and overlaid with PG&E plat maps to determine the features of each plat. Fractional values were achievable based on size and shape differences between county tracts and utility plats. For example, a census tract may contain several plats or portions within its boundaries. Once complete, scores between 1 and 10 were assigned to indicate feature criticality. The table below shows the assigned values aligned with scaled and normalized feature scores. Using the SUMPRODUCT function in Excel, each plat row was multiplied by the criticality array and summed, resulting in a unique Vs between 0 and 1.

Table 5.1: Example plat data with criticality matrix

	PLAT FEATURES WITH CRITICALITY WEIGHTS								
PLAT NUMBER	Hospital	Clinic	Government Services	Emergency Services	Grade School	University	Building Total	Population	Service Taps
2639-H08	0.0000	0.0663	0.0471	0.0014	0.0997	0.0000	0.2717	0.1799	0.4282
N/A	10	10	4	8	2	2	6	8	5

The Vs allowed us to differentiate between plat importance when examining the results of the model. By taking the average cost and multiplying the value by 1 plus the Vs we accounted for regions that are considered more critical for emergency response operations. For example, a class-1 plat with a Vs of 0.15 that was misidentified can be assigned a cost of $163,543 USD. This is shown in equation 11, where the initial cost is the result of the average incident value after the probability of occurrence is taken into consideration.

$$Plat\ Cost = \$142{,}211.5 \times (1 + Vs) \tag{11}$$

By taking the derived plat values and comparing them with the results of Table 3.4 model value can be quantitatively represented through cost savings. Values for plats that had leaks but were not identified can be summed to determine the total cost of incorrect predictions. Using the testing sample previously described from this historical dataset, improved predictive performance resulted in approximately 5.5 times less cost. Table 5.2 displays the results.

Table 5.2: Financial improvement from proposed model

COST SAVINGS (USD)	
TOTAL COST PROPOSED MODEL	$ 796,384.40
TOTAL COST PG&E MODEL	$ 4,373,003.63

5.2 Future Work

While the predictive performance of the model will be improved following the collection of new observations during the next major event, there are steps that can be taken immediately to improve upon the work.

Class imbalance concerns may be mitigated through techniques including over sampling, under sampling, and a combination of the two. Considering the relatively small size of the dataset, Synthetic Minority Over-Sampling Technique (SMOTE) could potentially improve the current results. In this method, decision boundaries along the minority class are relaxed as additional samples are generated. Additional samples are generated along line segments joining selected samples and their k-nearest neighbors. The difference between a selected feature vector and its nearest neighbor is then multiplied by a random number between 0 and 1 and added to the originally chosen feature. This generates an additional point between the two selected features [41]. This application was briefly considered during model development and used in several experimental cases but those cases have not been presented in this work. Future iterations may consider use of this technique after a more robust assessment.

Secondly, there are numerous techniques available for predictive analytics that have not been discussed in this paper. While it is not beneficial to provide an exhaustive list, we would like to bring attention to

the use of Poisson regression techniques. Dealing with relatively rare occurrences resulting in non-normally distributed and skewed leak datasets make Poisson regression a potentially attractive tool. The output would provide counts instead of probability of class ownership, which could be beneficial, but would also remove discretionary decision-making responsibilities from the response crew. For example, analyzing the probabilities of class ownership not only provides a prediction, but gives the decision-maker a sense of the prediction quality. Regardless, the technique has been used in weather forecasting applications along with traffic accidents and flood occurrence predictions [42, 43]. Applying such a method is worth pursuing and comparing results through future iterations.

Examining data from the 1994 Northridge Earthquake just outside of Los Angeles, California could provide additional observations for the model. The Southern California Gas Company was not contacted during this research but could be used to build more robust datasets. Other more recent events have occurred but have fortunately been off-shore (2010 Eureka) or in more remote regions were damage to the natural gas distribution system was not recorded (2003 San Simeon).

Outside of California, Alaska has recorded 3 events with magnitudes of 6.0 or greater within the last three years. These could provide valuable insights as well into features that may help predict damage to distribution lines. Internationally, recent data from Mexico and Japan could also lead to improvements. The issue with using information from outside PG&E's service territory and adjacent areas is that infrastructure changes (line material, housing construction, population density, etc.) may skew results and lead to observations that would normally not occur in the Northern and Central California Region.

Without other historical observations, live event data can also be incorporated into the model to build larger training sets in real-time during an earthquake. Currently, the proposed model does not receive feedback from the leak survey teams regarding the number of leaks discovered in each plat. The reporting system can be linked with the model to provide additional training points. Once enough points have been accumulated, the model can be re-trained using the new data. This idea was further explored using semi-supervised learning techniques. If live information from the leak survey teams could populate the training set with a small number of labeled samples, an algorithm can be used that takes advantage of the geometry of both labeled and unlabeled data [44]. Graphical methods and label propagation algorithms were not pursued in depth but are worth studying further for future model iterations.

Thirdly, HAZUS-MH predictive data was not incorporated into the final model because the results were not beneficial. It is suspected that user error contributed to this fact. Our efforts to replicate building damage results from the 2014 American Canyon FEMA assessment were unsuccessful, suggesting further training is warranted. Future iterations may benefit from the inclusion HAZUS-MH software.

Finally, based on model performance using CAT I and II data, the project did not incorporate CDI information into the final model. PG&E already uses USGS ShakeMap data, and so implementation would come through the same channels. For this report, the data was manually entered in order to evaluate the effectiveness. It is recommended to include this feature in future modifications. Crowdsourcing data can be an effective means to assessing the severity of the situation from the perspective of individuals in the damaged area and may provide more useful information than below-ground sensors. During the 2014 American Canyon earthquake, DYFI recorded 25,000 responses within one hour of the event, leading us to believe that actionable insights can be gained in near real-time through this system.

Appendix A

Emergency Event Roles

Distribution Integrity Management Program (DIMP):

Once the GEC is activated, DIMP personnel are recalled to San Ramon to provide direction to Leak Survey and field crews to mitigate damage to the distribution system. Within 60-90 minutes, a report from the Dynamic Automated Seismic Hazard (DASH) model is reviewed by the Geosciences Department and emailed to critical personnel. For a detailed explanation of the model please see Appendix B. The report indicates which plats are predicted to have suffered the most damage. The model uses an algorithm that takes data from USGS sensors, inventory (e.g. pipe length and material), and geological survey data (e.g. liquefaction/landslide/soil corrosiveness) to make this prediction. Each affected plat is assigned an Earthquake Prioritization (EP) value, with the highest score indicating the most damage. These scores are then sorted, reviewed by DIMP, and Leak Survey personnel are sent out to assess the damages.

Leak Survey has limited vehicle assets (Picarro trucks) that are used to help localize major leaks. Even with the use of these trucks, personnel are still required to survey the area on foot, with hand-held monitoring devices. Leaks are reported in "A-Forms," which appear in SAP and are refreshed every night. Phone calls and emails are also used to relay data back to the GEC. If major leaks are discovered, Leak Survey will notify the GEC and Maintenance and Construction (M&C) to mitigate the situation. They are not equipped or qualified to handle major leaks and will require additional crews or pipeline engineers (PLEs) to make assessments and perform major repairs. Information from the field is then considered and plat survey assignments are reassessed to reflect the situation.

Gas Service Representatives (GSR):

Concurrent to the tasking of Leak Survey personnel, PG&E maintains a call center to respond to customer needs. Important or timely calls are filtered through this center and sent to the Dispatch team, located in San Ramon. In addition to customer calls, dispatch can receive calls directly from police, fire, and other emergency response agencies. Based on the call, the appropriate work center is notified to respond. Routinely, Dispatch directly assigns personnel based on their knowledge of the individuals work schedule and proximity to the customer, without needing to rout tasking through the GSR supervisor. In large events, such as an earthquake, Dispatch can make the decision to instead route assets to the OEC and allow the Incident Commander (IC) to make the assignments.

Regardless of the situation, a GSR will be assigned because they are the quickest to respond. Even if it is known that the situation will require a M&C crew, a GSR will be notified and act as the first responder. They maintain their own vehicles and can rapidly deploy as needed. The GSR will be assigned a job through the Field Automation System (FAS), which is fed into SAP. If the leak is deemed hazardous and requires an immediate repair (Grade 1), Dispatch will use the Incident Management Tool (IMT) to log the event and assign a crew. It is important to note that as of March 1st, 2017 the IMT system is being replaced by the Emergency Management (EM) tool. This tool is designed to streamline the process and allow Dispatch and Gas Control (GC) to have access and awareness to the same information.

Once Dispatch assigns a representative or crew to respond to an event, they do not follow-up or track their progress. Information regarding the status of a response is either accessed through SAP/FAS or received via a phone call or email. As previously noted, information from Dispatch and the GSRs is present

on the DIMP emergency response check-list (described Chapter 3), but GSRs are not required or instructed to share their status or results of their response calls.

Appendix B

Dynamic Automated Seismic Hazard (DASH) Description

The first actions taken after an earthquake are directly informed by the Dynamic Automated Seismic Hazard (DASH) model. When internet is available the model results are distributed to emergency personnel after being reviewed by the Geosciences department. If unavailable, emergency responders refer to a series of designated scenarios, previously generated by the model, based on major fault lines and USGS predictions of earthquake magnitude and location likelihood. DASH provides two critical data points that help drive the emergency response: repair rates per 1000 feet of pipe and survey prioritization values per plat. The repair rates are calculated using the American Lifelines Alliances (ALA), empirical data-based methodology for determining water pipeline damage estimates. These equations are listed in Appendix C, along with those used by the FEMA developed software HAZUS-MH.

In PG&E's case, model parameters have been adjusted to reflect material differences and gravity-fed versus pressurized lines. Survey prioritization values, known as Earthquake Prioritization (EP) Values, assign a specific number to a geographic area impacted by the earthquake. The size of each area, known as a plat, is dependent on the number of assets involved and is used in daily operations to assign maintenance tasks along specified pipeline segments. For example, a plat in the center of San Francisco may only be 50x75 meters, but in a rural community it can be as large as 800x800 meters. All of PG&Es service territory is divided among approximately 22,000 individual plats. After an earthquake, an algorithm assigns each affected plat a score based on the values below:

EP plat sheet = ((LS value + Liq value + fault value)/3 + 1) x PGA

Where,

Fault value = 1.0 when fault lines (Historic/Holocene) cross plat sheet and

M = 6.0 or greater

Fault value = 0.0 if no fault crosses the plat sheet, or when fault lines

(Historic/Holocene) cross plat sheet with M < 6.0

PGA = peak ground acceleration in %g

LS value and Liq value from Table B.1 and represent landslide susceptibility (LS value) and liquefaction susceptibility (Liq value)

Leak survey teams use these initial EP values to determine where they send their assets. Any plat that receives a value greater than 40 must be surveyed [29]

Table B.1: Liquefaction and landslide values used by PG&E for determining EP values

Hazard	Liquefaction-Lateral Spreading Values			Landslide Values		
ShakeMap PGA(g)	Moderate to High	Low to Moderate	Low	Moderate to High & Known	Moderate	Low to Moderate
0.05	0	0	0	0	0	0
0.10	0	0	0	.2	0	0
0.15	.1	0	0	.4	0	0
0.20	.2	0	0	.6	0	0
0.25	.3	.1	0	.8	0	0
0.30	.4	.2	.1	1.0	.1	0
0.35	.6	.25	.15	1.0	.2	0
0.40	.8	.33	.2	1.0	.4	0
0.45	1.0	.4	.25	1.0	.6	0
0.50	1.0	.5	.3	1.0	.8	0
0.55	1.0	.6	.35	1.0	.9	0
0.60	1.0	.8	.4	1.0	1.0	.1
0.65	1.0	1.0	.47	1.0	1.0	.2
0.70	1.0	1.0	.53	1.0	1.0	.4
0.75	1.0	1.0	.6	1.0	1.0	.6
0.80	1.0	1.0	.7	1.0	1.0	.8
0.85	1.0	1.0	.8	1.0	1.0	.9
0.90	1.0	1.0	1.0	1.0	1.0	1.0
0.95	1.0	1.0	1.0	1.0	1.0	1.0
1.00	1.0	1.0	1.0	1.0	1.0	1.0

APPENDIX C

Damage Prediction Equations as Expressed Through Repair Rates

FEMA HAZUS-MH Model

Table C.1: Repair rate calculations used in HAZUS-MH software. Table modified directly from Table 8-21 in HAZUS-MH Technical Manual Version 2.0

	PGV Algorithm		PGD Algorithm	
	$R.R. \cong 0.0001 \times PGV^{(2.25)}$		$R.R. \cong Prob[liq] \times PGD^{(0.56)}$	
Pipe Type	Multiplier	Example	Multiplier	Example
Brittle Pipeline	1	Steel Pipe w/ gas weld	1	Steel Pipe w/ gas weld
Ductil Pipeline	0.3	Steel Pipe w/Arc weld	0.3	Steel Pipe w/Arc weld

Where,

R.R. = Number of repairs per kilometer

Prob[liq] = Conditional liquefaction probability relationships derived from type of deposit and sediment distribution. Tables are also available for public use. For purposes of this thesis they have not been copied over into this appendix.

PG&E Model

Table C.2: Repair rate scaling factors adapted to account for PG&E inventory and pipe pressurization (table taken directly from [3] without alterations)

	K_1			K_2		
Pipe Material	**Corrosive Soils**	**Non-Corrosive Soils**	**Unknown Soils**	**Corrosive Soils**	**Non-Corrosive Soils**	**Unknown Soils**
Cast Iron	1.4	0.7	1.0	1.4	0.7	1.0
Steel	0.9	0.3	0.6	0.9	0.3	0.6
Plastic (PE)	0.3			0.3		

(1) $R.R. = K_1 \times 0.00187 \times PGV$
Where,
R.R. = Estimated number of repairs per 1000 feet of pipe
K_1 = Constant scaling factor used to reflect differences in expected performance from the baseline pipe (small diameter cast iron pipe, with cement joints) resulting from ground shaking; determined by pipe material, joint type, diameter, and soil corrosivity
PGV = Peak ground velocity, in inches/second

(2) R.R. = K_2 x 1.06 x $PGD^{0.319}$
Where,
R.R. = Estimated number of repairs per 1000 feet of pipe
K_2 = Constant scaling factor used to reflect differences in expected performance from the baseline pipe (small diameter cast iron pipe, with cement joints) resulting from ground shaking; determined by pipe material, joint type, diameter, and soil corrosivity
PGD = Peak ground displacement, in inches

APPENDIX D

Dataset Features Listed by Category (CAT)

Table D.1: Model Features

Category	Features														
CAT I	EP	PGA	LIQ	LS	FAULT	PIPE_P	PIPE_S	CALLS							
CAT II	EP	PGA	LIQ	LS	FAULT	PIPE_P	PIPE_S	CALLS	TAPS	POP	BLDGS				
CAT III	EP	PGA	LIQ	LS	FAULT	PIPE_P	PIPE_S	CALLS	TAPS	POP	BLDGS	CDI	CDI RESP		
CAT IV	EP	PGA	LIQ	LS	FAULT	PIPE_P	PIPE_S	CALLS	TAPS	POP	BLDGS	CDI	CDI RESP	SURVEY DATE	PREV. LEAKS

Where,
EP = Earthquake Prioritization Value
PGA = Peak Ground Acceleration
LIQ = Liquefaction susceptibility value
LS = Landslide susceptibility value
PIPE_P = Length of plastic pipe (ft)
PIPE_S = Length of steel pipe (ft)
CALLS = Number of calls PG&E received from customers and emergency services personnel

TAPS = Number of service taps
POP = Population
BLDGS = Number of buildings
CDI = Community Decimal Intensity as calculated through the USGS 'Did You Feel It" (DYFI) program
SURVEY DATE = Number of days since plat was last surveyed
PREV. LEAKS = Previous number of leaks found in plat from most recent survey results

APPENDIX E

Notes from Author to PG&E Personnel Regarding Model Construction

The below information represents notes from the author to PG&E personnel describing features of the model. Color coded headers were in place to correlate with additional spreadsheets that were used while presenting model development efforts to management and engineering teams.

PLAT NUMBER
2639-H08

Description: PG&E separates their service territory into plats, based on the density of assets. For example, a plat in San Francisco may be 50m x 50m, but a plat in Napa could be an area closer to 600m x 800m. The number and letter combination are simply a naming mechanism used by the organization.

Author Notes: Data uploaded directly from PG&E mapping. The shapefile used is considered very accurate.

EP	PGA	LIQ_VALUE	LS_VALUE	FAULT_VALUE
32	0.235	0.3	0.8	0

Description: **EP** value refers to the "Earthquake Prioritization" value assigned through the DASH model. The EP value is the end product of an algorithm designed to predict pipeline damage based on the intensity of the earthquake and location and properties of the asset. The algorithm uses the peak ground acceleration (**PGA**) from USGS ShakeMaps, a liquefaction susceptibility value (**LIQ_VALUE**), a landslide susceptibility value (**LS_VALUE**), and a proximity to fault (**FAULT_VALUE**) to predict damage to the pipeline. The susceptibility values indicate whether or not the pipe is located in an area that is predisposed to liquefaction or landslides following a seismic event. Other factors are used in the algorithm, but the ones listed above are seen directly in the output generated by the DASH model.

Author Notes: The output is taken directly from DASH. Throughout an event, multiple iterations are produced based on USGS analysis and updated sensor information. The DASH output used in this model is version 19 of the 2014 American Canyon event and is considered the most accurate. The reason we chose to use such a late revision is because the Napa earthquake was a unique case and occurred along a previously unidentified fault. The model assumption is that a major event will occur along a known line, and the initial reports will be more accurate. However, it is also known that the liquefaction value can erroneously assign damage. During the Napa earthquake, liquefaction did not occur, and plats covering areas predicted to experience liquefaction were assigned EP values that did not reflect actual damage. This is a point of uncertainty within the model.

LEN_PIPE_P	LEN_PIPE_S	TOTAL LEN	% PLASTIC	%STEEL
1209	805	2014	60.02979146	39.97020854

Description: These values are an extension of the above elements and are also outputs of the DASH model. **LEN_PIPE_P** (ft.) denotes the total length of plastic pipe in the plat, **LEN_PIPE_S** (ft) denotes the total length of steel pipes in the plat, **TOTAL_LEN** (ft.) is the summation of the two lengths, and the percentage blocks refer to each material's respective contribution to the total amount of pipe in the plat. It is important to note that these values only refer to the distribution lines. There are currently no datasets available with to the total length of service and branch lines.

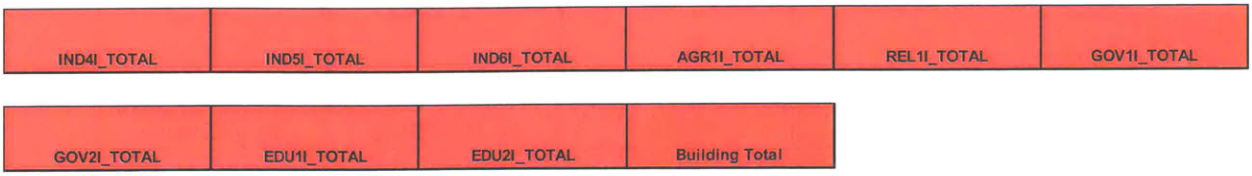
Author Notes: The data is a direct output from DASH and is considered accurate.

RED_BLDG	BLDG_NORMALIZED
4	0.01656692

Description: The **RED_BLDG** column refers to the number of buildings that were red-tagged as a result of the earthquake. Red tags indicate that a structure is not safe to occupy. Following an earthquake, multiple organizations and academic institutions conduct damage surveys to collect information on the impact of the event. The data for the Napa earthquake comes from the Earthquake Engineering Research Institute (EERI) reconnaissance survey. The goal of the model is to populate this column with predictive data from FEMAs HAZUS-MH model. The survey results are currently acting as a proxy for this data until it can be obtained. The **BLDG_NORMALIZED** is calculated by dividing the number of red-tagged buildings by the total number of buildings in the plat.

Author Notes: The data in this column was collected post-event and is therefore considered accurate. As mentioned above, this column could be populated from predictive analysis using the HAZUS-MH model. Literature regarding the accuracy of this model suggests it has a tendency to over-predict building damage. By using USGS data from the event and updating the building inventory to better reflect structures within PG&Es service territory we aim to mitigate the over-prediction.

RES1I_TOTAL	RES2I_TOTAL	RES3AI_TOTAL	RES3BI_TOTAL	RES3CI_TOTAL	RES3DI_TOTAL
RES3EI_TOTAL	RES3FI_TOTAL	RES4I_TOTAL	RES5I_TOTAL	RES6I_TOTAL	COM1I_TOTAL
COM2I_TOTAL	COM3I_TOTAL	COM4I_TOTAL	COM5I_TOTAL	COM6I_TOTAL	COM7I_TOTAL
COM8I_TOTAL	COM9I_TOTAL	COM10I_TOTAL	IND1I_TOTAL	IND2I_TOTAL	IND3I_TOTAL



Description: The codes represent building types as they are referenced in the FEMA damage prediction software HAZUS-MH. The tables below are extracted from the HAZUS User's Manual and provide a description for each code. FEMA lists the building types per census tract, which is why fractional numbers appear in the model. **Building Total** refers to the complete number of buildings estimated in each plat. This value is used to normalize the red-tagged building data.

RES1I	Single family dwellings
RES2I	Manufactured housing
RES3AI	Duplex - 1 to 2 units
RES3BI	Duplex - 3 to 4 units
RES3CI	Duplex - 5 to 9 units
RES3DI	Duplex - 10 to 19 units
RES3EI	Duplex - 20 to 49 units
RES3FI	Duplex - more than 50 units
RES4I	Temporary lodging
RES5I	Institutional dormitories
RES6I	Nursing homes
COM1I	Retail trade

COM2I	Wholesale trade
COM3I	Personal and repairs services
COM4I	Professional and technical services
COM5I	Banks
COM6I	Hospitals
COM7I	Medical office and clinic
COM8I	Entertainment and recreation
COM9I	Theaters
COM10I	Parking garages
IND1I	Heavy industrial
IND2I	Light industrial
IND3I	Food/drugs/chemicals

IND4I	Metal/minerals processing
IND5I	High technology
IND6I	Construction facilities and offices
AGR1I	Agriculture facilities and offices
REL1I	Churches and non-profit organizations
GOV1I	Government - general services
GOV2I	Government - emergency response
EDU1I	Grade schools and administrative offices
EDU2I	Colleges and universities

Author Notes: The number of building were taken from each tract and then assigned to individual plats. This was done in the same manner as the population calculation. The total number of buildings in the tract is multiplied by the percentage of area the individual plat covers. The sum of all these areas (depending on how many tracts an individual plat covers) is then taken. This data assumed building density and make-up are uniformly distributed within the tract. This will create additional uncertainty in the model.

CDI	NRESP	NRESP_NORMALIZED
7.6	3	0.055530258

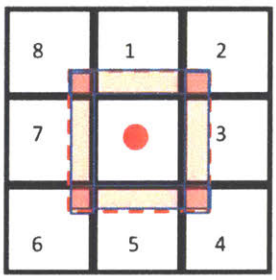
Description: The Community Decimal Intensity (**CDI**) is the result of an initiative from the USGS known as "Did you Feel it?" The data is collected from the general populace through an online survey. The survey answers are compiled and entered into an algorithm to determine the CDI. It can be thought of as another intensity measurement. In the Napa earthquake, over 40,000 people responded to the survey, most of which within the first two hours. If addresses are included, USGS creates a geocoded map to indicate where the CDI values originated. The size of the area is 1 km x 1 km and is formed using UTM coordinate boundaries. For this dataset, the centroid of the area was chosen and matched with a corresponding plat. The number of responses is indicated in the **NRESP** column. **NRESP_NORMALIZED** is the normalized value of the number of responses per plat population. It is the calculated by taking the number of responses and dividing by the plat population.

Author Notes: Responses are considered accurate based on the number of participants. Irregular responses are filtered out through the summation of total responses and further examined by USGS personnel. As mentioned above, the geocoded areas are larger than traditional plats and are based on the location of the responses. This means that aggregated responses, assigned to an individual plat, may have come from a neighboring area. Additionally, plat population was calculated using the tabular intersection tool in ArcGIS. PG&E does not maintain datasets with this information. In order to calculate the plat population, the plat map was overlaid on top of 2012 census tract data. The percentage of the tract that the plat occupies was then multiplied by the total tract population. This process was done until the entire plat area was accounted for. This method does well for population densities that are evenly distributed within the tract, but not for large, rural areas for population density is biased towards a specific area.

MAX CDI (Neighbor)	MAX NRESP (Neighbor)	N-NRESP_NORMALIZED
7.5	5	0.090954734

Description: The **MAX CDI (Neighbor)** is the largest CDI value attained by each plat, while it was acting in a neighbor capacity. This terminology is explained in the "Confidence" section below, but in brief, "neighbor" refers to any plat that is adjacent to the plat in consideration. The CDI value used for the model falls within a single plat (as a point value) but could have been calculated using data from adjacent plats because USGS and PG&E use different grid systems. The **MAX NRESP (Neighbor)** represent the maximum number of responses the plat received while acting in a neighbor capacity. The **N-RESP_NORMALIZED** is the value for the number of responses divided by the total population of the plat.

Author Notes: Based on the nature of the CDI geolocation methodology, it cannot be definitively determined if the CDI value came from a single plat. The UTM grid and plat maps are not the same size and do not overlap in the same manner. The CDI value used in the model is represented on the map as a point value, located at the centroid of the 1 KM by 1 KM UTM square. Therefore, it is conceivable that the responses, and subsequent CDI values were actually calculated from one of the plats eight neighbors. The graphic below shows and example:



The red circle represents the point value of the CDI. It falls within a single plat, each represented as a black square. The neighboring plats are numbered, which prevent plat neighbor "1" from having that specific identity multiple times. The red square shows the UTM box used by USGS. The shaded red parts show that while the CDI centroid falls within a single plat, the data used to calculate the CDI could have come from any of the neighboring plats. To account for this, each of the eight neighboring plats was assigned the same CDI value. Plats with missing data were then filled with the highest value they received while acting as a neighboring plat. Plats with CDI values already assigned by USGS were not altered, even if they received a higher value during the neighbor calculation. This was to maintain the integrity of the original data.

This is a very conservative approach to the problem, in which there are many solutions. An average of the plats CDI values while acting in a neighboring capacity could also be used to handle missing data points. The most conservative approach was chosen based on the severity of a potential gas leak. Over predicting the number of leaks is seen as a better result than failing to predict.

CALLS	POP	CALLS_NORMALIZED
6	1053.7515	0.005693942

Description: **CALLS** refer to calls received from dispatch. They are from customers and emergency response organizations (e.g. Police and Fire) requesting assistance from PG&E. **POP** refers to the population of the plat from which the calls came, and **CALLS_NORMALIZED** is the number of calls divided by the total population. By doing this we are weighting multiple calls from sparsely populated plats more heavily to increase the accuracy of the prediction.

Author Notes: Call data is taken directly from FAS. The raw feed has been scrubbed for the Napa region and is considered accurate. Population data is from the 2012 census and is likely to have increased in the last 5 years.

Days Since Survey (5 YR)	Days Since Survey (1 YR)
848	440

Description: **The Days Since Survey (5 YR)** column refers to the last time the plat was surveyed, under the five-year schedule. Notice that some plats extend beyond five years because they were delinquent in surveying. **The Days Since Survey (1 YR)** is the same measure, but for plats on a one-year inspection schedule. Some of these extend beyond 1 year for the same reason. It is not clear why the same plat may have the different service schedules. It appears that different assets may be checked at different times, but this has not been confirmed. More interviews need to be conducted to figure out why this occurs.

Author Notes: Data provided through DIMP (Distribution Integrity Management Program) and there is a high-level of confidence in the accuracy. It is unclear why data is missing for certain plats.

PREVIOUS LEAKS (2008-2013)	Grade 1	Grade 2	Grade 2+
40	2	35	3

Description: **PREVIOUS LEAKS (2008-2013)** represents the number of leaks found by the Leak Survey teams between the years 2008 and 2013. They have been further broken down into **Grade 1** (requiring an immediate repair), **Grade 2** (requiring repair within 18 months), and **Grade 2+** (requiring repair within 90 days). **Grade 2+ no longer exists at PG&E.**

Author Notes: Data provided through DIMP and there is a high-level of confidence in the accuracy.

2008 GR1	2009 GR1	2010 GR1	2011 GR1	2012 GR1	2013 GR1
2008 GR2	2009 GR2	2010 GR2	2011 GR2	2012 GR2	2013 GR2
2008 GR2+	2009 GR2+	2010 GR2+	2011 GR2+	2012 GR2+	2013 GR2+

Description: These columns indicate the year and leak grade that was found during leak survey inspections. **GR1** indicates a grade 1 leak, **GR2** indicates a grade 2 leak, and **GR2+** indicates a grade 2+ leak. The definitions for each are listed in the previous description. A number that is **bold and red** is a recorded leak that does not fit within the time of the last survey. This can be accounted for through call outs from GSRs or other employees that required a leak survey check.

Author Notes: Data provided through DIMP and there is a high-level of confidence in the accuracy. An issue that arose considered missing data points. It is unclear if the missing cells are meant to represent zero leaks or un-recorded information. The assumption made was that if any of the columns contained a data point, the missing cells in the row were filled with zeroes. These points are later refined after examining survey dates. Blank cells that are outside the leak survey dates have been labeled as "NA," and the data is considered non-available. If every column was blank, then it was assumed the data for that plat is missing. In this instance, "NA" has been recorded. On occasion, a leak will appear, but there will be no recorded leak survey within the year the leak was found. These data are kept because they could have resulted from a call-out from dispatch or M&C.

Description: The **SERVICE TAPS** are the number of meter connections within each plat. The meter sets are often a point of failure during a seismic event, and having this data was seen as potentially valuable for the model.

Author Notes: The dataset was compared with the Gas Distribution GIS portal, which is updated daily. A random sample of plats was compared between the two datasets to assess the accuracy of the set used for the model. The model set was consistent with the daily updated values. The difference was general on the order of +/- 7%. This level of error was seen as acceptable for use in the model. This will serve as a source of error in the model because the numbers reflect current service taps. We were unable to obtain the number of service taps present in August of 2014.

APPENDIX F

Example of Independent Variable Matrix with Plat Indices

PLAT NUMBER	EP	PGA	LIQ_VALUE	LS_VALUE	FAULT_VALUE	LEN_PIPE_P	LEN_PIPE_S	TOTAL LEN	% PLASTIC	%STEEL	RED_BLDG	BLDG_NORMALIZED	RES1I_TOTAL
2639-H08	20	0.203	0	0	0	1209	805	2014	60.02979146	39.97020854	0	0	126.5686945
2639-I08	21	0.218	0	0	0	293	237	530	55.28301887	44.71698113	0	0	107.18115
2639-J08	23	0.226	0.1	0	0	0	37	37	0	100	0	0	19.9735
2640-H01	21	0.203	0.2	0	0	336	118	454	74.00881057	25.99118943	0	0	60.3302
2640-I01	23	0.228	0.1	0	0	2	128	130	1.538461538	98.46153846	0	0	60.03
2640-J01	25	0.247	0.1	0	0	0	212	212	0	100	0	0	15.138
2640-J02	27	0.25	0.3	0	0	0	0	0	NA	NA	0	0	15.138
2708-I7	25	0.21	0	0.6	0	123	0	123	100	0	0	0	52.775

RES2I_TOTAL	RES3AI_TOTAL	RES3BI_TOTAL	RES3CI_TOTAL	RES3DI_TOTAL	RES3EI_TOTAL	RES3FI_TOTAL	RES4I_TOTAL	RES5I_TOTAL	RES6I_TOTAL	COM1I_TOTAL	COM2I_TOTAL	COM3I_TOTAL	COM4I_TOTAL
50.1595968	2.8843245	2.08	0.8	0.16	0	0	1.6043245	4.8008649	0	1.6164331	0.8095139	1.2895139	2.5764331
43.14256	2.39015	1.716	0.66	0.132	0	0	1.33415	3.96283	0	1.37377	0.69113	1.08713	2.16577
11.3982	0.1815	0.091	0.035	0.007	0	0	0.1255	0.2211	0	0.2809	0.1571	0.1781	0.3229
26.0017	1.2104	0.8489	0.3265	0.0653	0	0	0.688	1.966	0	0.786	0.4035	0.5994	1.1778
38.272	0.23	0	0	0	0	0	0.23	0.046	0	0.874	0.506	0.506	0.874
9.6512	0.058	0	0	0	0	0	0.058	0.0116	0	0.2204	0.1276	0.1276	0.2204
9.6512	0.058	0	0	0	0	0	0.058	0.0116	0	0.2204	0.1276	0.1276	0.2204
3.2	2.6	0.4	0.125	0.25	0.025	0	0.35	0.125	0.025	0.7	0.475	0.625	1.125

COM5I_TOTAL	COM6I_TOTAL	COM7I_TOTAL	COM8I_TOTAL	COM9I_TOTAL	COM10I_TOTAL	IND1I_TOTAL	IND2I_TOTAL	IND3I_TOTAL	IND4I_TOTAL	IND5I_TOTAL	IND6I_TOTAL	AGR1I_TOTAL	REL1I_TOTAL
0.16	0	0.48	2.5695139	0	0	0.48	0.32	0.9712437	0	0	0.6460543	1.1303788	1.1269192
0.132	0	0.396	2.14313	0	0	0.396	0.264	0.82879	0	0	0.54781	0.95796	0.94664
0.007	0	0.021	0.2341	0	0	0.021	0.014	0.1863	0	0	0.1057	0.1822	0.1378
0.0653	0	0.1959	1.1218	0	0	0.1959	0.1306	0.4828	0	0	0.3102	0.5411	0.5131
0	0	0	0.506	0	0	0	0	0.598	0	0	0.322	0.552	0.368
0	0	0	0.1276	0	0	0	0	0.1508	0	0	0.0812	0.1392	0.0928
0	0	0	0.1276	0	0	0	0	0.1508	0	0	0.0812	0.1392	0.0928
0.1	0.025	0.4	0.925	0	0	0.125	0.15	0.25	0	0	0.425	0.35	0.3

GOV1I_TOTAL	GOV2I_TOTAL	EDU1I_TOTAL	EDU2I_TOTAL	Building Total	MAX CDI (Neighbor)	MAX NRESP (Neighbor)	N-NRESP_NORMALIZED	CDI	NRESP	NRESP_NORMALIZED	CALLS	POPULATION	CALLS_NORMALIZED
0.32	0.0008649	0.3260543	0	203.8807283	0	0	0	0	0	0	0	473.3787611	0
0.264	0.00283	0.28381	0	172.99961	0	0	0	0	0	0	0	400.56737	0
0.014	0.0111	0.0917	0	33.9977	0	0	0	0	0	0	0	73.1339	0
0.1306	0.007	0.1796	0	98.2776	0	0	0	0	0	0	0	224.6979	0
0	0.046	0.322	0	104.282	0	0	0	0	0	0	0	151.172	0
0	0.0116	0.0812	0	26.2972	0	0	0	0	0	0	0	54.9724	0
0	0.0116	0.0812	0	26.2972	0	0	0	0	0	0	0	54.9724	0
0.025	0	0.15	0	66.025	0	0	0	0	0	0	0	127.575	0

Days Since Survey (5 YR)	Days Since Survey (1 YR)	PREVIOUS LEAKS (2008-2013)	Grade 1	Grade 2	Grade 2+	2008 GR1	2009 GR1	2010 GR1	2011 GR1	2012 GR1	2013 GR1	2008 GR2	2009 GR2
848	440	40	2	35	3	0	1	0	0	0	1	0	26
844	439	10	0	7	3	0	0	0	0	0	0	0	4
846	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
853	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
834	440	3	1	2	0	0	1	0	0	0	0	0	1
846	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

2010 GR2	2011 GR2	2012 GR2	2013 GR2	2008 GR2+	2009 GR2+	2010 GR2+	2011 GR2+	2012 GR2+	2013 GR2+	SERVICE TAPS	FAS LEAKS	IMT LEAKS	LS LEAKS	TOTAL	Classifier
2	1	5	1	0	2	0	0	0	1	352	0	0	0	0	0
2	0	1	0	0	1	1	0	1	0	29	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	80	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	4	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	18	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	0	0	0	0	0

* Not all of the variables were used in model construction but were compiled into a single dataset for later segregation. Leak values were documented in green and separated based on their source. Once confirmed, the number of leaks were summed and used to create the appropriate classifier. These values would then be removed from the matrix, leaving only the predictor variables and corresponding classifier for training the model.

Bibliography

[1] Pacific Gas and Electric Company. (2018) *Company Profile* [Online]. Available: https://www.pge.com/en_US/about-pge/company-information/profile/profile.page

[2] California Public Utilities Commission. (2017) *Natural Gas and California* [Online] Available: http://www.cpuc.ca.gov/general.aspx?id=4802

[3] H. Seligson, "Technical Summary of Pipeline Repair Estimation Methodology and Application for PG&E Natural Gas Transmission and Distribution Pipelines," MMI Engineering Project No. MMHB065. Tech. Memo. MMHB065, Oct. 06, 2014.

[4] Southern Gas Association Gas Measurement Training Council. (2011). *Industry Overview* [Online] Available via company intranet: http://pgeweb/gas/standards/integrity/fimp/gasmeasurement/_layouts/15/WopiFrame.aspx?sourcedoc=/gas/standards/integrity/fimp/gasmeasurement/Documents/Training%20Docs/TA2-SBWBT-IndustryOverview.pptx&action=default&DefaultItemOpen=1

[5] Working Group on California Earthquake Probabilities, "USGS Earthquake Probabilities in the San Francisco Bay Region: 2002-2031," United States Geological Survey. Open-File Report 03-214. 2003.

[6] PHMSA. (2018). *Pipeline Incident Database* [Online]. Available: https://hip.phmsa.dot.gov/analyticsSOAP/saw.dll?Portalpages&NQUser=PDM_WEB_USER&NQPassword=Public_Web_User1&PortalPath=%2Fshared%2FPDM%20Public%20Website%2F_portal%2FSC%20Incident%20Trend&Page=All%20Reported&Action=Navigate&col1=%22PHP%20-%20Geo%20Location%22.%22State%20Name%22&val1=%22%22

[7] G. Avalos (2017, Apr. 21). *Court OKs $90 million PG&E San Bruno Explosion Settlement* [Online]. Available: https://www.mercurynews.com/2017/04/21/court-oks-90-million-pge-san-bruno-explosion-settlement/

[8] Federal Emergency Management Agency. (2017). *Hazus Success Stories* [Online]. Available: https://www.fema.gov/hazus-success-stories

[9] Federal Emergency Management Agency, "Multi-Hazard Loss Estimation Methodology," Mitigation Division, Technical Manual. Hazus-MH 2.1, Jul 26, 2013.

[10] Utility Procedure: TD-4110P-09 - Leak Grading and Response, Rev 5, Pacific Gas & Electric Company, 2016.

[11] American Lifelines Alliance, "Seismic Fragility Formulations for Water Systems, Part 1 – Guideline," FEMA and ASCE, Contract Report. Apr. 2001.

[12] D. G. Honegger, "Repair Patterns for the Gas-Distribution System in San Francisco", U.S. Geological Survey Professional Paper 1552-A, Aug. 05, 1997.

[13] J.K. Virostek, S.H. Phillips, "Natural Gas Disaster Planning and Recovery: Loma Prieta Earthquake," Pacific Gas and Electric Company, Company Report. Apr 1990.

[14] "Improving Natural Gas Safety in Earthquakes," Prepared by ASCE-25 Task Committee on Earthquake Safety Issues for Gas Systems, California Seismic Safety Commission, Jul 11, 2002.

[15] H. Seligson, "Review of the Distribution Pipeline Leak Data from the 2014 Napa Earthquake," MMI Engineering Project No. MMHB072. Tech. Memo. MMHB072-TM-002, Rev 0., Aug. 24, 2015.

[16] H. Seligson, "Meter Set Modeling Exercise, DASH Damage Model Technical Support," MMI Engineering Project No. MMHB072. Tech Memo. MMHB072-TM-001, Rev 0, Jun 09, 2015.

[17] A. Domijan Jr, A. Islam, W.S. Wilcox, R.K. Matavalam, J.R. Diaz, L. Davis, and J. D'Agostini "Modeling the Effect of Weather Parameters on Power Distribution Interruptions," in *7th IASTED Int. Conf. Power and Energy Systems, Clearwater Beach, Fl, USA, 2004*.

[18] G. D. Eschelbach, "Wires-Down Predictive Modeling and Preventative Measures Optimization", Master's Thesis, Massachusetts Institute of Technology, 2016.

[19] H. A. Nefeslioglu, E. Sezer, C. Gokceoglu, A. S. Bozkir, and T. Y. Duman, "Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey," Mathematical Problems in Engineering, vol. 2010, Article ID 901095, 15 pages, 2010.

[20] C. Ballabio, S. Sterlacchini, "Support Vector Machines for Landslide Susceptibility Mapping: The Sraffora River Basin Case Study, Italy," *Mathematical Geosciences*, vol. 44, issue 1, pp. 47-70, Jan. 2012.

[21] B. Pham, B. Pradhan, D. Tien Bui, I. Prakash, M.B. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environmental Modeling and Software*, vol. 84, pp. 240-250, Oct. 2016.

[22] J. Bialas, "Object-based classification of earthquake damage from high-resolution optical imagery using machine learning", Master's Thesis, Michigan Technological University, 2015.

[23] T. Sakaki, O. Makoto, and M. Yutaka, "Earthquakes Shakes Twitter Users: Real-Time Event Detection by Social Sensor," *Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, April 26, 2010*. pp. 851-860.

[24] C. Boulton, H. Shotton, H.T.P. Williams, "Using Social media to Detect and Locate Wildfires," in *AAAI Publications, Tenth International Conference on Web and Social Media, 2016*. pp. 178-185.

[25] Cabot Oil & Gas Corporation. (2014, Jun. 02). *Natural Gas Distribution Expanding in Northeast PA* [Online]. Available: https://wellsaidcabot.com/natural-gas-distribution-expanding-in-northeast-pa/

[26] J. Lemus, "Issue Investigation Procedure Napa Earthquake DIMP Mitigation," Pacific Gas & Electric Company, Tech, Report. 2014-20, Oct. 2015.

[27] L. Krefta, "Asset Knowledge & Integrity Management Earthquake Playbook," Pacific Gas and Electric Company, Gas Operations, Asset Knowledge and Integrity Management, Instruction, Attachment 6. Nov 05, 2015.

[28] United States Geological Survey. (2017). *Did You Feel It?* [Online]. Available https://earthquake.usgs.gov/data/dyfi/

[29] L. Krefta, "Asset Knowledge & Integrity Management Earthquake Playbook," Pacific Gas and Electric Company, Gas Operations, Asset Knowledge and Integrity Management, Instruction. Nov 05, 2015.

[30] S. Raschka. (2014). *About Feature Scaling and Normalization* [Online]. Available: http://sebastianraschka.com/Articles/2014_about_feature_scaling.html#about-standardization

[31] G. Bautista, M. Monrad, "A Study of K-nearest Neighbor as an Imputation Method," In *Soft Computing Systems – Design Management and* Applications, HIS, 2002, Santiago, Chile, December 1-4, 2002.

[32] Scikit-Learn Developers. (2017). *Ensemble Methods* [Online]. Available: http://scikit-learn.org/stable/modules/ensemble.html#forest

[33] S. Raschka, *Python Machine Learning*. Birmingham, United Kingdom: Packt Publishing Ltd., 2015.

[34] Scikit-Learn Developers. (2017). *Precision-Recall* [Online]. Available: via http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[35] United States Geological Survey. (2017). *ShakeMap* [Online]. Available: https://earthquake.usgs.gov/data/shakemap/

[36] H. Fujiyoshi, Y. Mishina, R. Murata, T. Yamashita, Y. Yanauchi, "Efficient Feature Selection Method using Contribution Ratio by Random Forest," in *Frontiers of Computer Vision (FCV), 21st Korea-Japan Joint Workshop, 2015, Mokpo, South Kora, May 11, 2015.*

[37] M. Kurdi, "Optimizing Emergency Response Crew Allocation During Earthquakes to Improve Restoration Efforts," Master's Thesis, Massachusetts Institute of Technology, 2017.

[38] C.G. Rubeiz, "Performance of Pipes During Earthquakes," *in Pipelines Specialty Conference 2009, San Diego, CA, USA, August 15-19, 2009.*

[39] Pipeline and Hazardous Materials Safety Administration. (2018). *PHMSA's Mission* [Online]. Available: website https://www.phmsa.dot.gov/about-phmsa/phmsas-mission

[40] PHMSA. (2018). *Pipeline Incident 20 Year Trends* [Online]. Available: https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-20-year-trends

[41] K. Bowyer, N. Chawla, L. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, pp.321-357, 2002.

[42] L. Thakali, K. Kanitpong, and M. Hossain, "Development of Accident Prediction Models, their Possibilities and Efficacy for the Developing Countries; a Thai Experience," in *Proceedings for the Eastern Asia Society for Transportation Studies, Vol 7, 2009.*

[43] M. Cupal, M. Deev, abd D. Linnertova, "The Poisson Regression Analysis for Occurrence of Floods," in *2nd Global Conference on Business, Economics, Management and Tourism, 30-31 October 2014, Prague, Czech Republic.*

[44] Y. Bengio, O. Delalleau, and N. Le Roux, "Label Propagation and Quadratic Criterion" In *Semi-Supervised Learning*, O. Chapelle, B. Scholkopf, and A. Zien, Eds. Cambridge, MA: MIT Press, 2006, pp. 193-216.