SMU Data Science Review

Volume 2 | Number 2 Article 16

2019

Machine Learning in Support of Electric Distribution Asset Failure Prediction

Robert D. Flamenbaum
Southern Methodist University, rflamenbaum@smu.edu

Thomas Pompo
Southern Methodist University, tpompo@mail.smu.edu

Christopher Havenstein
Southern Methodist University, chavenstein@mail.smu.edu

Jade Thiemsuwan
Southern Methodist University, jthiemsuwan@semprautilities.com

Machine Learning in Support of Electric Distribution Asset Failure Prediction

Robert D. Flamenbaum, Thomas Pompo, Christopher Havenstein, Jade Thiemsuwan

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{rflamenbaum, tpompo, chavenstein}@smu.edu
jthiemsuwan@sdge.com

Abstract. In this paper, we present novel approaches to predicting asset failure in the electric distribution system. Failures in overhead power lines and their associated equipment in particular, pose significant financial and environmental threats to electric utilities. Electric device failure furthermore poses a burden on customers and can pose serious risk to life and livelihood. Working with asset data acquired from an electric utility in Southern California, and incorporating environmental and geospatial data from around the region, we applied a Random Forest methodology to predict which overhead distribution lines are most vulnerable to failure. Our results provide evidence that a predictive model can be built with the data at hand, but policies such as purging failed asset records are problematic for producing highly predictive models that can be used for proactive asset management.

1 Introduction

Electric utilities are an important part of modern society's infrastructure, supplying 4,178 billion kilowatt hours of electricity in 2017 and serving more than 150 million customers in the United States alone [1]. While technological improvements in electric distribution devices continue to improve power delivery and quality, the basic structure of the system remains largely unchanged over the past century. As such, electric infrastructure can be decades old in many neighborhoods and in need of repair or modernization. Premature aging of electric facilities can result from adverse environmental conditions and configuration maladies, leading to increased susceptibility to failure. Factors such as geography, weather, and wire size are among a host of variables that must be considered when evaluating and managing asset health [2]. Given the complexity of the issue, machine learning techniques are ideal for explaining the uncertainty that confounds traditional asset health management models, thus increasing reliability and preventing unplanned outages.

By producing a machine learning model for predicting failure of overhead power-lines, our results can potentially be used to increase reliability, as well as

reduce financial, regulatory, and environmental risk for these utilities. This is especially relevant in California where the combination of dry heat, high wind speeds, and electrical device malfunction can lead to a serious and persistent threat of wildfire. By improving the reliability of the electric utility, there are various benefits to be gained. Not only does a reliable company appear attractive to investors, but there is additional benefit of reaping performance incentives from regulatory authorities as opposed to paying fines for failure to meet reliability goals. Furthermore, as the threat of wildfire diminishes, so does the threat of resulting lawsuits.

Our asset data comes from a Southern California utility and includes information on the asset itself as well as all failures which have occurred in the region since 1981. The asset data is combined with geospatial, environmental and weather data to add new features and increase the predictive capabilities of our model. Cleaning and preparing the data was necessary to ensure that the data from our various sources were aligned correctly in order to give accurate results. Furthermore, important variables such as Asset Age needed to be imputed due to missing data. Asset age is particularly important as it serves as a baseline for determining asset health in conventional asset management practices. Given the importance of this variable to the model, it was necessary to depart from simple imputation models, such as using the median age for all assets, and instead employ a Random Forest classification algorithm over 25 independent variables.

After identifying our most important features for classifying age of the asset based on data exploration and visualization, as well as domain knowledge from experts in the field, we were left with a complete data set of overhead power-lines in major population areas of Southern California. With this data set, we compared various machine learning techniques including Logistic Regression and Random Forest, with the latter having the best results. We furthermore evaluated Synthetic Minority Over-sampling Technique (SMOTE) and Random Under-sampling to remedy unbalanced data sets [3] [4].

We were able to create an age classification model that predicts asset age with 82% accuracy. The age is broken into 10 year bins according to decade from 1960 to 2019. Assets older than 1960 are grouped together as any asset older than 60 years is considered beyond its service lifespan. Our precision and recall averaged over each age bracket are each 82%. The resulting classifier was applied to the set of population data that contained NULL work order dates. The result of this imputation was a data set with zero NULL values for chronological age.

The imputed asset age variable was subsequently added to a data set that was used to classify the population of overhead conductors as outages vs non-outages. The results of this model were that we were able to predict 63% of outages, with an AUC (Area Under the Curve) of 68%.

The results of our work show two things. First, that asset age can be reliably imputed using a Random Forest classification algorithm over variables that include asset, geographic, and environmental data. This imputation method should prove useful for any future use case where work history information is lacking.

And second, given our current data and the limitations we faced, we can build a somewhat predictive model, but not highly predictive.

The remainder of this paper is organized as follows. In Section 2 of our paper we present more background and tutorial information on electric utilities including history, measures of reliability, a look at current mitigation practices, and the benefits of an updated process. In Section 3 we delve further into the data, including our sources, the data collection process, and a closer look at the data preparation. In Section 4 we walk through our methodology and the building of our model. In Section 5 we examine our results while in Section 6 we analyze the results. In Section 7 we discuss the ethical implications of our research and the electrical utility industry as a whole. In Section 8 we deliver our conclusions and our suggestions for future research.

2 Electric Distribution: A History and Overview

Electric power in the United States has existed for well over 100 years. The first power grid came online in San Francisco in 1879, followed by the Niagara hydro-electric plant in 1896 [5]. Since these pioneering efforts, electric power in the United States has grown into an asset that has largely changed the ways in which the the nation functions. It has also helped to propel the United States into a role as a leading power in the world.

The electric power grid consists of three major components: generation, transmission, and distribution. Electricity on the electrical grid originates with power generation. Electric generation plants produce electricity by transforming energy into electricity using a variety of methods including thermal energy, such as is produced by fossil fuel plants; and potential energy, which is exemplified by hydroelectric power plants. While fossil fuel and hydro electric plants have produced the bulk supply of electricity for decades, wind and solar electric generating facilities are becoming more common as the demand for renewable, clean energy sources increases. Recently, the state of California has set a goal of obtaining 60% of the state's total energy from renewable resources by 2030, ultimately reaching 100% renewable energy production by 2045. Once electricity is produced, it is available to be purchased by electric utilities and scheduled to enter the transmission grid.

The electric transmission system is designed to carry high voltage electricity over long distances. Electricity in the electric transmission system can travel through multiple utility jurisdictions, state boundaries and even national boundaries. Transmission voltages typically range anywhere between 500 kV and 69 kV. The equipment used in the transmission system consists of large robust structures, wires, and devices that are designed to handle the high voltages that flow through the electric transmission system.

From the transmission grid, electricity flows to substations where the power is stepped down to primary distribution voltages, which typically range from 12 kV to 4 kV. From a substation, power enters the primary electric distribution system where it travels along overhead or underground circuits. The primary voltage can

either be stepped down to secondary electric distribution voltages, or carried directly to customers. Electricity from the primary or secondary distribution system is subsequently passed through transformers where it is further stepped down to 240/120 volts, making it consumable via customer alternating current (AC) outlets.

When considering power outage mitigation, there are two main categories to distinguish between; transmission and distribution. Although transmission networks are connected to distribution networks, the systems are often modeled independently of each other. Transmission networks have separate maintenance schedules as well as components that are distinct to transmission voltages [6]. While outages in the electric transmission system can be severe, these events affect customers to a lesser degree than distribution outage events [2].

For the purposes of this study, only outages in the electric distribution system are used in the analysis. Distribution electric has a high degree of complexity due to the wide diversity of components used in the the system [7]. Distribution electric also has the added element of chaos associated with it due to the high number of un-monitored energy consumers tied to the network. Whereas most of the transmission outage issues are known to be associated with inclement weather, machine learning has the potential to solve the riddles of the more complex distribution system. Distribution outages pose a great amount of risk when considering the myriad of critical energy usage implementations such as life support devices, air conditioning, heating, and road infrastructure [7]. In the context of public safety, power outages can lead to increased crime. Also of concern is the issue of food borne illnesses as detailed in the article, "Food safety during power outages," [8]. Distribution outages do not only pose a risk to life and limb, but are also used in the indices that regulating authorities use to gauge an electric utility's reliability [9].

2.1 Basic Electric Distribution System Protection

Throughout the electric distribution system there are built in protection mechanisms that help ensure as few customers as possible are impacted by faults. The first means of protection for a circuit after leaving a substation is the circuit breaker. A circuit breaker will open, or cut off the flow of electricity when fault current exceeds specified normal operating parameters. From the circuit breaker, electricity travels along a mainline, which is comprised of a thick gauge wire, typically larger than #2 size wire. The mainline is also referred to as the backbone of the circuit. Throughout the mainline, there are protection devices such as fuses, switches, and dynamic protection devices installed. Similar to a circuit breaker, these devices are engineered to open when current exceeds specified operating conditions. From the mainline, the circuit branches off into many areas in order to deliver electricity to customers. Located at the beginning of each of the branches, wire size usually will transition from a large gauge to a smaller gauge wire. The hardware used to secure the wire will also change according to design standards. Depending on the count and location of customers, additional protection devices are installed to further mitigate customer outage

impact. For the utility on which this study is based, it is important to note that design specifications for the electric distribution system are continually maintained and updated by the electric distribution engineering standards group, as new knowledge is gained from investigating past faults.

2.2 Reliability and the Importance of Mitigation

Reliability is a key factor in determining an electric utility's success[10]. Reliability is based upon both the number of outages that a utility experiences and the length of time that a customer is without power. The more reliable service that an electric utility provides, the more attractive it appears to investors. Likewise, reliability is monitored by federal and state regulatory authorities where rewards or penalties can be inflicted depending on the utility's ability to improve upon their reliability numbers [6] [9]. In the state of California, the California Public Utilities Commission (CPUC), is responsible for rewarding utilities for improved reliability scores The CPUC is furthermore responsible for penalizing utilities for poor reliability metrics. On an annual basis, the utility upon which this study was conducted is subject to rewards or penalties that could amount to approximately $4,000,000. The main indices used to gauge reliability for the utility in question are the System Average Interruption Duration Index (SAIDI), and the System Average Interruption Frequency Index (SAIFI)[9]. SAIDI measures the outage duration that a customer experiences in units of minutes per year [11]. The SAIDI target for the utility in question is 62 minutes per year. In other words, on average, a customer should experience little more than one hour of unplanned service interruptions in the year. SAIFI measures the average quantity of outages experienced by a customer per year [11].

In addition to reliability, safety issues abound when considering electric distribution infrastructure. An inherent risk of electrical power is fire. Notable of late, are the wildfires attributed to electric infrastructure in Northern California. Though investigations are still pending, Pacific Gas and Electric (PG&E) has been the subject of intense scrutiny for unintentionally starting wildfires via their electric facilities. As such, PG&E recently filed for chapter 11 bankruptcy protection to stem the enormous payouts anticipated from litigation over the death and damage resulting from fires in 2017 and 2018. Because electric utilities are being held responsible for fires originating from their facilities, regardless of whether their facilities were determined to be out of compliance, it is a paramount policy to mitigate potentially faulty devices lest a fire develop from a resulting fault. Recently, San Diego Gas and Electric (SDG&E), a utility which has experienced large wildfires in the past two decades, unveiled a plan for wildfire mitigation that involves the large scale deployment of synchrophasers [12]. Synchrophasers have the ability to detect fallen conductors quickly enough to turn off the power prior to igniting a wildfire during adverse weather conditions [12]. In addition to the death and destruction resulting from wildfires caused by electric infrastructure, subsequent lawsuits have the ability to destroy electric utility companies, leaving citizens at a loss for clean, reliable electric power.

There are many asset management methodologies that could be implemented for SAIDI and SAIFI mitigation. While these practices differ in some ways, they typically center around how much of the system to restore and when [2]. Common to all of these methodologies is the need to mitigate the unknown factors that cause reliability numbers to increase [13]. The unknown factors that contribute to SAIDI, SAIFI, and safety risk are precisely what this paper seeks to remedy through the use of implementing a machine learning model on a comprehensive data set of the assets that comprise the electric distribution system.

2.3 Current Practices

The CPUC has a set mandated inspection cycle for poles, conductors, and cables. In accordance, utilities must conduct detailed inspections of conductor and cable within both urban and rural areas on a 5 year basis [14]. Patrol intervals for equipment in "Extreme and Very High Fire Threat Zones" is one year [14]. When an inspection finds an asset that is out of compliance, it has a finite time period to remedy the problem based on three levels of severity. General Order 95 Rule 18A states that Level 1 violations, or hazards that pose an "Immediate safety and/or reliability risk with high probability for significant impact," must be fixed within 6 months [15]. For example, if 100 wire spans are known to be Level 1 hazards, those 100 spans must be fixed within 6 months or the company faces stiff fines and penalties. The inspection cycle is designed to give the utility a reasonable amount of time to fix problems found during an inspection year.

The electric utility for which the data for this study was obtained operates in a fashion typical of those within the jurisdiction of the California Public Utilities Commission (CPUC). While there are many models for proactive asset replacement, preventive maintenance is ultimately tied to budgets and sanctioned projects [13]. The bottom line in any proactive maintenance project is to determine the most efficient way to protect the public and the company from risk due to asset failure. Some budgets will center on fire risk mitigation, such as the SDG&E synchrophaser implementation described above [12]. For a project such as this, the utility will propose a budget whereby the CPUC will approve, deny, or alter the proposal. Once a budget is settled, the issue of how many devices can be installed for the budget is determined by engineering and field personnel. After the amount of device installations is agreed upon, the question of where and when to schedule construction begins. Traditionally, expert opinion is used to assess where along the electric distribution system construction should occur. To maximize budget and construction efficiency, circuits are ranked according to the severity of risk. In the case of synchrophaser placement, public spaces such as schools and parks take top priority. For example, the likelihood of a passerby being injured from a falling conductor is greater than that of a vacant lot for instance.

A common practice for reducing SAIDI outage duration is to add sectionalizing devices that can finely isolate outage areas so that as many customers as possible can remain energized while the failed device undergoes repairs [10]. Placement of sectionalizing devices will typically be chosen by an experienced

engineer, who will determine placement based on reducing the maximum amount of impact on customer outage time. A unique consequence of operating an electric utility in drought stricken California is that reliability savings that can be achieved with the implementation of automatic reclosing sectionalizing devices are often forgone because of the fire risk associated with automatic reclosing of switches. If an energized line has fallen on the ground, the auto reclosing procedure could produce sparks from the fallen conductor, and thus start a potentially catastrophic wildfire. Risk in the basic context of likelihood of an event occurring versus the impact of the event is taken into account for any proactive replacement project [7].

The downside to traditional preventive maintenance practices is the difficulty involved with gauging the success of the replacement methodology. The only true measure of whether the current mitigation practices work is to compare reliability numbers from one year to another. This practice is imperfect at best as there are too many variables at play to clearly distinguish whether targeting a particular device for replacement is having an effect on reliability numbers. As such, the machine learning model created for this project seeks to mitigate the uncertainty and pitfalls associated with over-reliance on expert opinion for preventive maintenance.

3 Data Collection and Preparation

3.1 Data Sources

In order to create a comprehensive predictive model, data was collected from a variety of sources. Our primary data source is from a Southern California Electric Utility. This includes data on the asset itself, such as the type of asset, asset age, material, circuit, and length of cable. This also includes data on all of the failures that have occurred on these assets since 1981. We then used GIS (Geographic Information Systems) mapping technology to add geographic info such as elevation, slope, aspect and miles from the coast. Many enterprise databases were examined for usable data. While there were many interesting data sets available, they were often incomplete or out of date. To ensure timely relevance of our model, only regularly maintained data was incorporated into our data set for analysis.

3.2 The life of an electrical device through data

The primary objective for the data collection effort was to obtain a comprehensive picture of the life of the electrical devices. By understanding the life of an asset, we may be able to determine the conditions that lead to its death. Similar to predicting human lifespans based on demographics and lifestyle, it is theorized that different factors such as geography and configuration will play a role on the longevity of a deployed electrical device. While much is known about the electrical devices in the utility for whom this study was conducted, the data

is stored in many different databases where keys are inconsistent and data consistency is lacking. In order to piece together a complete story of the assets, an examination of the data collection methods and motivations is in order.

As the utility is well over 100 years old, the data collected over the years exists in varying conditions. The most apparent deficit for determining asset lifespan is the paucity of available installation dates. While this information does exist in hard copy and as image files in document management systems, very little is readily accessible via database. As such, chronological age of assets must be imputed using more consistent proxies.

The available asset data, which includes circuit, structure and device specific information, was extracted from the Electric GIS Production database. Asset data is created and maintained by a staff of GIS technicians, where as-built construction drawings serve as the source documentation for populating the database. This data was originally coalesced into a database during the 1990's as part of an Automated Mapping Facilities Management (AM/FM) project. This data was subsequently converted to a Geographic Information Systems (GIS) platform in 2011, where the data was normalized and network connectivity was applied. While the GIS conversion project vastly improved the geographic analysis capabilities of the asset data, a negative side-effect resulted in that much attribute data was lost, including installation dates. Asset data that exists in relatively complete states includes information such as wire sizes, pole material, transformer Kilovolt-amps (kVA), and connector types. The attributes collected as part of the predictive modeling effort will be detailed in an upcoming section.

An important dimension of the data that exists in a much more complete state centers around outages. The earliest available outage comes from 1981. This data set began its life as an ad-hoc project by engineers whose objective was to eventually be able to analyze the data for more adequately planning proactive maintenance projects. While the outage data from the 1980's is far from complete, records from the 1990's until the present are much more robust, due in large part to CPUC mandates centered around reliability [14]. Outage data includes the circuit effected, outage cause, damaged device, date of occurrence, and outage duration. This data is managed and scrutinized by an engineering team dedicated to reporting reliability information to regulating authorities as well as investors.

While asset data adequately describes the physical characteristics of electrical devices, it does not contain many variables that describe the environmental condition of assets' location. One way to accommodate missing environmental variables is to spatially derive the information using GIS. By using common GIS spatial analysis techniques, a myriad of variables can be extracted that describe the physical, environmental, jurisdictional, and demographic properties of the assets. For the purposes of this project, the variables extracted using GIS include elevation, aspect, slope, wind gust, lightning frequency, tree density, distance from the coast, and angle of orientation for the span.

Data retention policies played a crucial role in our ability to develop a highly predictive model. Past and current policies mandate that asset data is deleted

when a device is replaced in the field. As such, important attributes that contain the physical characteristics of failed devices are purged from the system of record and are likewise not archived. While we have confirmed that electric devices are replaced with similar devices, we cannot verify the exact configuration and model of the device that failed. For instance, a small wire gauge will be replaced with a similar small wire gauge, but we cannot verify the exact size, model, or material of the wire that failed. A #6 gauge wire is likely to be replaced with a slightly larger #2 size wire as #6 wire is being phased out of the system. Likewise, construction standards dictate that copper wire is to be replaced with aluminum wire. Therefore, it is impossible to identify finite problematic wire configurations as the data is not available. We can only make generalizations as to the wire size, and other characteristics of the failed devices.

3.3 Data Set Attributes

A total of 26 attributes were used in the Asset Age model. The overview of these attributes is shown in Table 1. The attributes include a unique identifiers (e.g., `feederid` and `conductorid`), physical attributes of the equipment and span (e.g., `wirematerial` and `measuredlength`), operating attributes of the equipment and span (e.g., `nominalvoltage` and `subtypecd`), and physical attributes of the installation (e.g., `elevation` and `treedensity`). A heat map showing the correlation of the attributes used in the Asset Age model is shown in Figure 1.

3.4 Preparing the Data

Data preparation was essential to the success of generating a comprehensive data set for asset age prediction. Based on domain expertise, Asset Age was expected to be an important variable in our outage analysis. Because installation of new assets has until recently been tracked in hard copy work orders, finding installation information is mostly a manual process that involves searching document management systems and hard copy documents. Some asset installation date information is available in a database format, but much of it is missing and must be imputed prior to using it in a model.

In order to impute asset installation dates, we used a classification algorithm to categorize the assets into 7 bins according to logical time groupings. The bin size is based on domain expertise regarding the quantification of electric device age. While engineers might prefer to know the exact year that an asset was installed, it would be difficult for an algorithm to accurately predict an exact year over the 60 plus year time frame. On the other hand, binning the assets into 20 year time frames might yield good predictability results, but would be too vague of a time span for proactively replacing aged assets. Working with subject matter experts, we chose the optimal time span in terms of both model predictability and actionable results. It was determined that 10 year increments would be sufficiently informative for asset managers to ascertain a basic age assessment. Therefore the decade variable was split into 6 categories accounting for each decade greater than or equal to 1960. All years prior to 1960 were included in

Table 1. Attributes of the data set used in the Chronological Asset Age Imputation model

Order	Field	Type	Source	Description
1	feederid	object	GIS	ID of circuit
2	measuredlength	float64	GIS	Length of span per as-built plans
3	conductorid	object	GIS	ID of span
4	subtypecd	category	GIS	Conductor Phase
5	nominalvoltage	category	GIS	Circuit voltage
6	backboneidc	category	GIS	Mainline or branch indicator
7	faultprotectiontype	category	GIS	Type of sectionalizing device on span
8	outage	int64	SAIDIDAT	Indicates if an outage occurred on span
9	wirematerial	category	GIS	Copper or aluminum wire
10	pole_wo_year	float64	GIS	Pole installation or refurbishment year
11	milestocoast	float64	GIS Derived	Distance in miles from ocean
12	elevation	float64	GIS Derived	Elevation of span
13	lightningdensity	int64	GIS Derived	lightning strikes per mile grid for span
14	windgust	category	GIS Derived	Expected wind gust for span
15	angleorientation	float64	GIS Derived	Circular angle of the span
16	treedensity	float64	GIS Derived	trees per mile for span
17	gauge	category	GIS Derived	large or small gauge wire
18	aspect	float64	GIS Derived	Direction of land slant for span
19	slope	float64	GIS Derived	Degree of land slope for span
20	decade	category	GIS Derived	Derived decade of span installation
21	pole_install_year	float64	GIS	Original install date of upstream pole
22	jointuseidc	category	GIS	Indicates non-electric utilities co-located
23	polematerial	category	GIS	Material of upstream pole
24	transmissionidc	category	GIS	Indicates transmission co-located
25	phasedesignation	category	GIS	Indicates phase of the span
26	stubidc	category	GIS	Indicates presence of stub pole

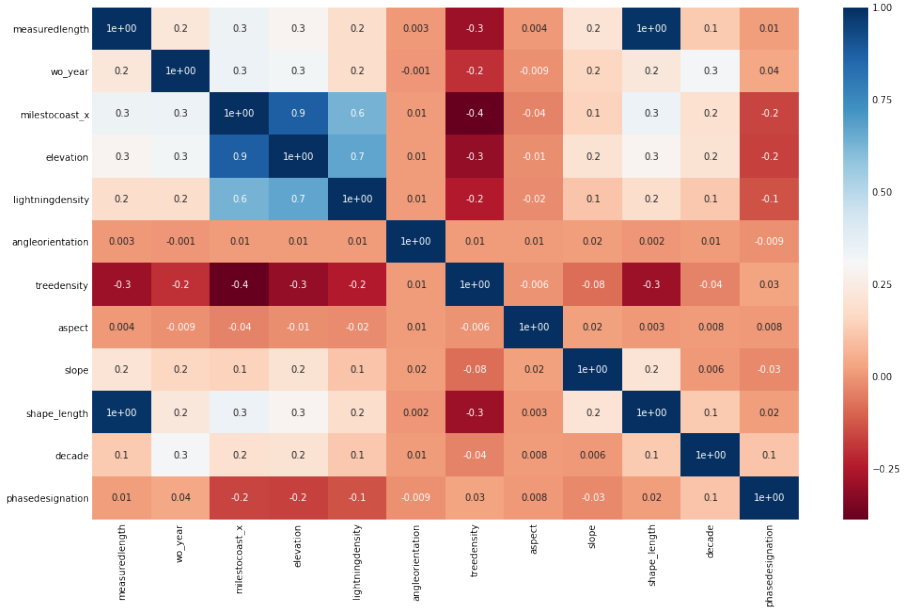

Fig. 1. A heat map showing the correlation of the independent variables in the asset age model

one category as all electrical equipment older than 60 years is considered past its lifespan. For overhead conductors, roughly 37,000 work orders exist for a population of roughly 170,000 spans.

3.5 GIS Variable Extract

The power of GIS data is that variables can be generated using spatial overlays and manipulations. Whereas the asset data did not have a variable for proximity of each asset to the coast, it was possible to extract this data by running a process to determine the distance in miles from each asset point location to the coast line. The ability to create variables in this fashion is important because domain experts believe that devices closer to the ocean will corrode much faster than the same types of devices situated further inland. There is in fact a GIS layer in the electric production database that demarks a contamination zone, or boundary where corrosion is expected to be prevalent on metallic surfaces. Using this same methodology variables were extracted for lightning density, elevation, average wind gust, directional angle of a span, and tree density. The inclusion of this spatially derived data with the asset data, dramatically fills in much of the unknown conditions that affect an electrical device's lifespan.

3.6 Creating the Data Set

In order to join the asset data, outage information, a common key needed to be created between the data sets. For the outage data, the circuit and the structure of the upstream outage device were concatenated together then joined with the same key formatted for the GIS data using the circuit and upstream structure variables. The resulting data set consisted of 48 variables. To create the outage response variable, every record that contained valid outage information was attributed as a 1. The records without outage information were attributed as 0's. While the outage records contained many descriptive variables surrounding the circumstances of the outage occurrence, these all had to be dropped from the data set because there were no corresponding variables for non-outage spans. These variables included information on the date and time of the outage, as well as the cause category and type of device that was damaged. This data set was ultimately pared down to 25 variables, and subsequently used in the Asset Age Imputation model.

3.7 Predicting Chronological Asset Age

The first step in generating predictions for asset age was to generate two data sets by separating the records with known work order dates from the records with NULL work order dates. The data set with the known work order dates was then split into train and test sets using sklearn's train test split functionality, with 70% of the data used for training and 30% for testing. The next step in the process was to use one-hot encoding to transform the categorical variables from the training set into a format the machine learning algorithm could use better in prediction. We subsequently normalized the data using scaling functionality from sklearn, which transformed the variables into a common scale.

To classify which decade group each span belonged to, we evaluated 2 classification algorithms. The first algorithm we used was K Nearest Neighbors (KNN) with 3 neighbors. The results for this algorithm did not show much accuracy with a score of 63%. While the accuracy may be improved by employing Grid Search to tune the hyper-parameters, we decided to evaluate a Random Forest algorithm on the data set. The results of this model, which are available in Table 4, showed substantially improved predictability with weighted averages of both precision and recall of 82%.

While chronological asset age imputation contained much value as a standalone use case, the primary use of the model for this project was to populate the missing ages for the records with missing work orders. The results of the Random Forest model show that the classifier has both strong precision and recall from 1990 until the present, then starts dropping in recall steadily from the 1980's and earlier. This coincides with the number of samples available in the data set, which can be seen in Table 3. Having substantially larger class sizes for the decades 1990, 2000, and 2010, suggests that the model is not as predictive in the earlier decades because of the sample size difference.

Table 2. Training data - decade frequency after cleaning data

Decade	Frequency
2010	9059
2000	10127
1990	10025
1980	1141
1970	148
1960	28
1950	74

Table 3. Classification report for Asset Age Imputation model

Decade	Precision	Recall	F1-Score	Support
1950	1.00	0.09	0.17	22
1960	0.00	0.00	0.00	8
1970	0.90	0.20	0.33	45
1980	0.85	0.31	0.46	400
1990	0.72	0.77	0.74	3008
2000	0.73	0.77	0.75	3038
2010	0.87	0.84	0.86	2718
accuracy			0.77	9181
Macro avg	0.72	0.43	0.47	9181
Weighted avg	0.77	0.77	0.76	9181

Once the classifier was trained and tested, it was applied to the data set with NULL work order dates to generate predictions. Subsequently, we applied the classifier to the entire data set and populated a new column with the predicted decade, which was used as an explanatory variable in the the Outage Prediction model. The addition of the new column furthermore allowed us to manually compare the actual work order dates for populated data with the predicted work order dates.

4 Building the Predictive Model

There were many caveats to consider when creating the outage prediction model. The first issue that needed to be resolved was the prediction units. For the purposes of overhead distribution outages, we settled on the span as our unit of measure. A span consists of a single overhead circuit from pole to pole. All of the wires, connectors, and devices are considered part of the span. The second major issue we addressed was the limitation of outage type scope. There was much debate as to whether the project scope should be limited to outages classified as equipment failure, thereby eliminating weather, customer contact, and crew error related outages from the list of failures. We decided to keep all overhead outages in scope based on the premise that even though inclement weather, mylar balloons, car crashes, and crew mishaps contribute to outages, there is always a device on the span that fails. Furthermore, limiting scope to just equipment failures would reduce the number of positive outages in our data set and cause our class imbalances to increase. Therefore, our scope includes all overhead outages in the electric distribution system.

4.1 Outage Prediction

With the completion of the decade imputation, a variable called decade was added to the analytics data set. In order to predict outages, the Outage binary column was set as the dependent variable. Working through the same methodology, the data was segregated into training and test sets using a 70% to 30% split ratio. We then implemented one-hot encoding for the categorical variables and scaled the data. To carry out this model, we used the Python scikit-learn library's Logistic Regression and Random Forest functionality.

The most notable characteristic of this data set was the large class imbalance in the outage variable. The non-outages accounted for 161,019 records whereas the outages accounted for 2,886 records. Ignoring the class imbalance to start, the data set was trained on a Random Forest classification algorithm. As can be seen in Table 4, the results of this initial run showed that the classifier was returning accuracy of 98%, which seemed to indicate that the class imbalance was causing the classifier to overly select the majority class. This notion is further supported in the classification report, that shows the recall for the minority class to be 13% recall, or the ratio of true positives over true positives plus false

positives, suggests that the classifier leaned towards classifying data in favor of the majority class.

In order to rectify the class imbalance we applied the Synthetic Minority Over-sampling Technique (SMOTE) method from the Imbalanced-Learn Python package to the training data. SMOTE works by synthesizing new data points based on inferences made on the configuration of the minority class data [3]. Using SMOTE, we were able to synthesize enough data so that both classes in the dependent variable were equal in number.

As an alternative to SMOTE, we utilized a Random Under-sampling method in which samples are randomly removed from the majority class to achieve balanced classes [4]. Although this method removes many of the cases, it has the benefit of increased precision over SMOTE.

After our classes were sufficiently balanced, we first implemented scikit-learn's Grid Search on a Logistic Regression algorithm using 5 fold cross validation. Using the best parameters as determined by the Grid Search process, the Logistic Regression model was run. In addition to Logistic Regression, we also utilized a Random Forest model, and compared the results of the two techniques.

5 Results

5.1 Outage Prediction Results

The results from the Logistic Regression model using SMOTE resulted in an Area Under the Curve (AUC) score of 62%. Most notable in the results of this model are that the recall for positively identified outages increased from 13% to 61%. This score indicates that SMOTE was able to substantially increase the model's ability to predict true positives and further shows the value of applying the up-sampling technique on our imbalanced data set. The precision score for class 1 is 3%, while the precision for class 0 or non-outages is 99%. The recall for class 0 is 58%. The full results of the unbalanced outage classification model and the SMOTE corrected model are available in Tables 4 and 5 respectively.

The Random Forest model using the Random Under-sampling method produced the most accurate results with a precision and recall scores of 65% and 63% respectively. The micro, macro, and weighted average scores were 63% across the board for precision, recall, and f1-score.

6 Analysis

Our results demonstrate that asset age imputation using a Random Forest algorithm is plausible. While the model had a weighted precision and recall of 77%, it was weak at predicting minority classes. The minority classes for this use case represent the older wire spans, which on a conventional level, are considered the most risky. Whereas the aim of this model is to feed the decade predictions into

Table 4. Classification report for Overhead Span Outage Prediction model without correcting for class imbalance

Value	Precision	Recall	F1-Score	Support
0	0.98	1	0.99	49393
1	0.87	0.13	0.23	950
micro avg	0.98	0.98	0.98	50343
macro avg	0.93	0.57	0.61	50343
weighted avg	0.98	0.98	0.98	50343

Table 5. Classification report for Overhead Span Outage Prediction model using Imbalanced Learn - Synthetic Minority Over-Sampling Technique (SMOTE)

Value	Precision	Recall	F1-Score	Support
0	0.99	0.58	0.73	50019
1	0.03	0.61	0.05	962
accuracy			0.58	50981
macro avg	0.51	0.59	0.39	50981
weighted avg	0.97	0.58	0.72	50981

Table 6. Classification report for Overhead Span Outage Prediction model using Random Under-sampling technique

Value	Precision	Recall	F1-Score	Support
0	0.62	0.64	0.63	930
1	0.65	0.63	0.64	975
macro avg	0.63	0.63	0.63	1905
weighted avg	0.63	0.63	0.63	1905

Table 7. Features ranked in terms of importance

Rank	Variable	Importance
0	milestocoast_x	0.105599
1	measuredlength	0.105154
2	angleorientation	0.10488
3	elevation	0.10345
4	treedensity	0.098049
5	aspect	0.097667
6	slope	0.09627
7	lightningdensity	0.072192
8	phasedesignation	0.026048
9	decade_pred	0.022446
10	wirematerial_CU	0.01551
11	jointuseidc_Y	0.013103
12	faultprotectiontype_F	0.01281
13	gauge_small	0.011569
14	faultprotectiontype_N	0.010838
15	elevationbin_101-500	0.010135
16	faultprotectiontype_R	0.009074
17	nominalvoltage_12.0	0.00877
18	stubidc_Y	0.008473
19	windgust_85.0	0.008367
20	elevationbin_501-1000	0.00831
21	polematerial_WOOD	0.008179
22	subtypecd_3	0.006965
23	subtypecd_2	0.005767
24	elevationbin_1001-2000	0.005661
25	backboneidc_Y	0.005579
26	polematerial_WEATH	0.005482
27	faultprotectiontype_E	0.004319
28	elevationbin_2000+	0.003968
29	polematerial_STEEL	0.003694
30	windgust_111.0	0.001617
31	transmissionidc_Y	0.000055

Table 8. Features ordered from left to right as seen in Figure 3

Order	Variable
0	measuredlength
1	milestocoast_x
2	elevation
3	lightningdensity
4	angleorientation
5	treedensity
6	aspect
7	slope
8	phasedesignation
9	decade_pred
10	subtypecd_2
11	subtypecd_3
12	nominalvoltage_12.0
13	backboneidc_Y
14	faultprotectiontype_E
15	faultprotectiontype_F
16	faultprotectiontype_N
17	faultprotectiontype_R
18	wirematerial_CU
19	windgust_85.0
20	windgust_111.0
21	gauge_small
22	elevationbin_1001-2000
23	elevationbin_101-500
24	elevationbin_2000+
25	elevationbin_501-1000
26	jointuseidc_Y
27	polematerial_STEEL
28	polematerial_WEATH
29	polematerial_WOOD
30	transmissionidc_Y
31	stubidc_Y

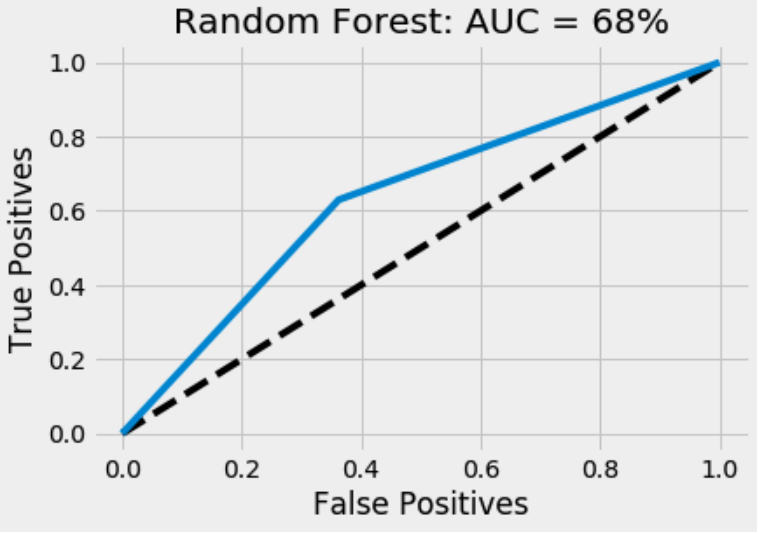

Fig. 2. A Graph showing the AUC of 68%

the Outage Prediction model, the overall accuracy score is acceptable. If however the the use case was to identify the oldest wire spans in an effort to replace them, the model would be insufficient. There two distinct groups that can be identified in the data by the number of samples with accompanying precision and recall scores. The decades, 1990's, 2000's, and 2010's all have sample sizes around 3,000 records. In turn, they all have relatively high recall scores ranging in the 77 to 84. On the other hand, the older decades (1980's and earlier), which have substantially less samples of 400 or less, display much lower recall scores at 31 or less. The correlation between sample size and higher recall scores suggests that the ability to predict older assets might be improved if additional samples can be generated through the research of work order records.

Table 8 shows the features used in our Random Forest model, ranked by importance. Our decade prediction variable, while in the top 10, was not as important as we expected, but asset age would likely be of more value if the age records were complete or the age predictions were more precise. Many of our most valuable features were the GIS derived ones, including elevation, miles to the coast and angle of orientation. This speaks to the value of interdisciplinary methods, this model far outproducing a model based solely on the asset data.

The Outage Prediction model using SMOTE resulted in a recall rate of identifying true positives of 61 %. While this result is substantially improved over the initial 13% recall we received without balancing the classes, the precision is very poor at 3%. The result is that our model is predicting too many false positives. These scores are unacceptable for productionizing the model within the utility. Budgetary limitations make it impractical to use the model for scheduling

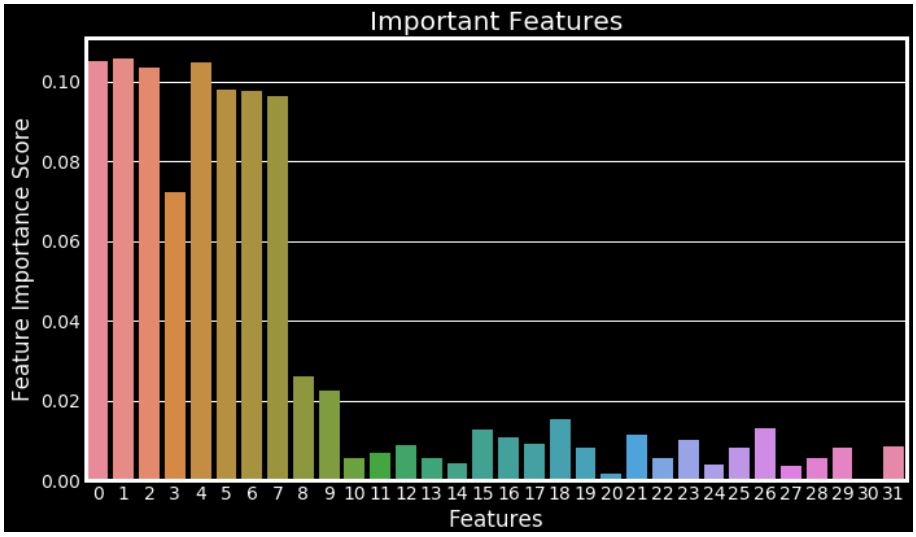

Fig. 3. Feature importance plot. Refer to Table 8 for variable order.

construction projects. In order to realistically consider this model for implementation in the real world, the precision needs to be drastically improved.

The best outage prediction score was produced by using the Random Under-sampling method. Using this method, the majority class was sampled so that it was equal in number to that of the minority class. In the case of our model, the majority class was reduced from 161,019 to 2,886. The resulting Random Forest model produced an area under the curve (AUC) of 68%, which was 6% higher than the AUC produced using the SMOTE method. Furthermore, the precision score produced by the model was 65%, which demonstrates the model's ability to distinguish true positives from false positives is vastly improved over the model using the SMOTE sampling technique. The drastic increase in precision and the moderate increase in recall using the Random Under-sampling method indicates that the model may be quite effective for risk mitigation. Considering the cost of construction, the tested ability to predict outages makes the productionizing of this model feasible.

7 The Burden of Knowing

A serious ethical issue exists when considering the appropriate response that is triggered when an electrical asset is found to be out of compliance. As stated earlier, spans that are known to be Level 1 hazards, must be fixed within 6 months. While it is advantageous for a utility to use predictive modelling for proactive asset replacement, the time restriction on fixing level 1 hazards makes fixing potential hazards cost prohibitive and logistically impossible. This dilemma has

a direct bearing on predictive analytics projects such as the work presented in this paper. When a proven predictive model determines an asset is in danger of failing, an off-cycle inspection will be triggered. If the device is found to be faulty, a work order will be issued and work will be scheduled to remedy the problem. While this scenario is well within the maintenance capabilities of a utility, there are plausible situations that pose a major risk for the company. Whereas the prediction of a single faulty asset poses no serious logistical maintenance issues, the implications of a model predicting the imminent failure of 1,000, 10,000, or even 100,000 assets is far different.

More serious implications would take effect if a wildfire or other disaster was caused by a device that failed an inspection. Fixing 100,000 wire spans within a 6 month time period is an insurmountable feet for even the most efficiently run utilities. Work order processes take time and collaboration by many departments to ensure construction standards are followed and quality workmanship is carried out. The work order process for such a hypothetical situation starts with obtaining an emergency budget for the project. Second, a skilled workforce required for the effort must be mobilized and trained. When considering the workforce required for the task, not only would a host of linemen be necessary for the task, but an ample amount of designers, mappers, land managers, environmental specialists, cultural resource managers, and many other specialists would be required to make sure the jobs are completed correctly. Needless to say, a utility would not want to be in a situation where it had to fix 100,000 assets within a 6 month time period. The preceding situation is why the mandated inspection interval is designed to balance safety within logistical capabilities.

The question of whether highly predictive machine learning model results are tantamount to physical inspections needs to be addressed. Currently there are no formalized protocols from regulating authorities that dictate the proper response for analytics results that indicate possible asset failure. Utilities must decide the point at which analytics results require action. There are currently no standards in place that dictate when a predictive model is accurate enough to constitute an inspection on positive asset failure results. While a model with an AUC of 75%, might not necessitate remedial action, it is possible that no remedial action taken for a model with an AUC of 95% would constitute negligence on the part of the utility. Furthermore, a device failure resulting in death or wildfire, that was predicted by the model to fail, could result in devastating settlement losses for the company. On one hand, utilities are motivated to engage in developing predictive models for proactive maintenance, however, there is a catch in knowing that problems exist, which causes some in the industry to shy away from engaging in predictive analytics.

The ethical and legal implications of predicting electrical device failure are complex. Regulating authorities must work with utilities to develop protocols for addressing predicted compliance issues within the maintenance capabilities of the utility. While the complexities of predictive analytics make it difficult to determine appropriate actionable standards, it stands to benefit both the utility

and customer to proactively seek out problem devices though the adoption of machine learning and statistical modelling.

8 Conclusions and Future Work

We have found that we can accurately predict and impute age for our assets with missing data. This is already valuable in and of itself, as there are numerous assets with unknown age due to poor record keeping by electric utilities. Refining the asset imputation model so that it more accurately identifies older assets and field verification of the predictions will further increase the ability for this model to be used for asset management. Research on hard copy work orders to increase the sample sizes of the older assets will furthermore increase the model's ability to correctly classify spans installed prior to 1990.

The asset failure prediction model, based on a Random Forest classifier, is able to identify 63% of failures, which makes productionizing the model feasible for construction planning and risk mitigation purposes. The results of this study provide a basis for identifying overhead spans in danger of failing. Considering the 65% precision and 63% recall for the positive outage records, the utility could reasonably scope out construction projects and be assured that their expenditures will mitigate outages 63% of the time.

The prediction capabilities of this model could be vastly improved by implementing a data retention policy where failed assets are not purged from the system of record. While we have been successful at developing an asset failure predictive model, data retention policies, or the lack thereof, have inhibited our ability to form a descriptive data set that contains reliable asset ages and configurations for failed electrical devices. Refining the data set through work order research, coupled with an in-depth study on failed asset configuration will help further increase the model's ability to predict true positives. Immediate reconsideration of the current data retention policy would help to ensure that future predictive modelling efforts will display increasing accuracy over time.

One way to increase the utility of the model is to look at the data from a daily health perspective. Daily high, low and average voltage data can be taken from SCADA to create a data set that continuously measures the health of the asset to better predict when an outage may be on the horizon. Furthermore, incorporating daily weather data rather than average weather data, could increase the predictive capabilities of the model as well. By recognizing the warning signs on a daily basis, resources can be better configured to prevent failure.

While we are looking at a broad range of failures in our model, we are only looking at a specific type of asset, those being overhead power-lines and their associated components. Given the importance of these types of assets, and the amount or risk and damage that goes along with them, it was decided to focus on these aspects. Expansion of the asset types in the model is the next step, whether that be in one overall asset failure model, or several specific models.

This is an observational study and our data looks specifically at assets in Southern California for one electric utility. Thus, our results are only applicable

to this region and this company. Expansion of the data to include other regions and other utilities would prove beneficial in widening the scope of utility for the model.

References

1. U.S. Energy Information Administration: Electric power annual 2017 - revised (December 2018)
2. Amin Moradkhani[1], Mahmood R. Haghifam[1], M.M.: Failure rate modelling of electric distribution overhead lines considering preventive maintenance. IET Generation, Transmission & Distribution **8** (June 2014) 1028–1038(10)
3. Chawla, N.V.e.a.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research (2002) 321–357
4. Zhang, C.X., Wang, G.W., Zhang, J.S., Guo, G., Ying, Q.Y.: Irusrt: A novel imbalanced learning technique by combining inverse random under sampling and random tree. Communications in Statistics - Simulation and Computation **43**(10) (2014) 2714–2731
5. Howe, C.: Power to the people. Science **353**(6297) (2016) 355–355
6. Heo, J., Kim, M., Lyu, J.: Implementation of reliability-centered maintenance for transmission components using particle swarm optimization. International Journal of Electrical Power & Energy Systems **55** (2014) 238 – 245
7. Carnero, M.C., Gomez, A.: Maintenance strategy selection in electric power distribution systems. Energy **129** (2017) 255 – 272
8. Gupta, R., Douglas, J.: Food safety during power outages. West Virginia Medical Journal, vol. 111, no. 5, 2015, p. 50+. Academic OneFile (2019)
9. Wang, B., Camacho, J.A., Pulliam, G.M., Etemadi, A.H., Deghanian, P.: New reward and penalty scheme for electric distribution utilities employing load-based reliability indices. IET Generation, Transmission & Distribution **12** (August 2018) 3647–3654(7)
10. Ferreira, G., Bretas, A.: A nonlinear binary programming model for electric distribution systems reliability optimization. International Journal of Electrical Power & Energy Systems **43**(1) (2012) 384 – 392
11. Kornatka, M.: Selected indicators of the national distribution system dependability. Acta Energetica **nr 4** (2013) 27–36
12. Fairley, P.: Utilities roll out real-time grid controls: Synchrophasor tech enables rapid response to broken power lines and other emergencies - [news]. IEEE Spectrum **55**(10) (Oct 2018) 9–10
13. Selvik, J., Aven, T.: A framework for reliability and risk centered maintenance. Reliability Engineering & System Safety **96**(2) (2011) 324 – 331
14. Public Utilities Commission of the State of California: General order number 165 (1997 updated 2009 2012)
15. Public Utilities Commission of the State of California: General order number 95 (2009 updated 2012 2017)