1

Predicting Low Voltage Events on Rural Micro-Grids in Tanzania

Samuel Steyer, Shea Hughes, Natasha Whitney

***Abstract*—Our team initially set out to predict when and where low voltage events and subsequent outages would occur on micro-grids in rural Tanzania. Our first hypothesis was that weather parameters would drive most of the system variations. To test this hypothesis, we compiled a feature space composed of grid statistics and weather data for a year, and began implementing machine learning algorithms including logistic regression, support vector machines (SVM), principle component analysis, and random forests to make our predictions. We were able to predict outages with a 20 % generalization error twelve hours into the future using random forests, and our original hypothesis was proven false. Our model suggests that the presence of an outage in the future is not directly correlated with any one feature; it is due to the highly stochastic state of the system at any given time.**

I. Introduction

OVER 1.2 billion people do not have access to electricity around the world today, and the International Energy Agency estimates that global energy demand will grow 36% by 2035. Developing countries will account for 93% of this increase; yet only 40% of the demand is likely to be met by grid-electricity [1].

The remaining 60% of electricity demand will be met by some combination of off-grid, micro-grid, or stand-alone systems. In recent years the reliable and scalable operation of micro-grids (small scale power grids that can operate independently) has arguably been the fastest changing, most dynamic aspect of the global energy system.

Our partner in Tanzania, Devergy, informed our group that one of the biggest challenges to maintaining a rural micro-grid is managing brown-outs and black-outs. These systems are often very remote and difficult to reach, and providing proper system management and maintenance requires careful planning. Through trial and error, Devergy discovered that quality of service drives micro-grid success. With poor service, customers grow impatient and usage drops.

Devergy monitors power generation and consumption using smart meters to see when voltage drops slowly over the course of several hours. To prevent outages, Devergy can either limit total household consumption, or install new solar photovoltaic (PV) generation towers close to customers using the most energy.

For our project, we construct a model to predict outages on rural micro-grids that serve agricultural communities so that grid operators (DESCOs) such Devergy can anticipate outages and maintain customer trust.

De Pascale, Fabio. Devergy.

Through discussions with Devergy and supplementary research, we learned that weather patterns and seasonal changes typically drive the majority of load variability and probability of grid failures. We therefore begin by exploring how best to apply learning algorithms to predict low voltage events and corresponding system outages based on a set of weather parameters and other features described in more detail below.

II. Data Processing

Devergy provided us with microgrid data for two villages in rural Tanzania, Doma and Mlandizi. The micro-grids for each village are laid out as a series of interconnected nodes. Each node is either a 60 Watt solar PV tower referred to as an enbox (represented by a sqaure), or a smart meter (represented by a circle). See Fig. 1. The grid grows in an organic manner as new homes become connected and new supply comes online over time.

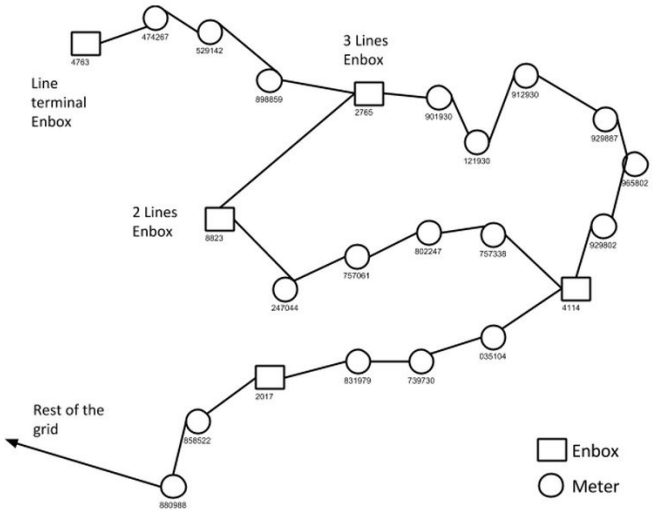

Fig. 1. Schematic of part of Mlandizi microgrid

The data provided by Devergy included current, voltage in, voltage out, and nodal 'mode' readings for all meters and enboxes (approximately 300 per grid) at irregularly spaced intervals over a year.

We used the 'voltage in' readings at the meters to spot low voltage events on the grid. The meters are configured to run household appliances at 24 V but have some built-in flexibility. According to Devergy, performance of the devices attached to the meters begins to suffer at voltage levels less than 18 V

and the coincident occurence of low voltage readings across multiple meters leads to a grid-wide outage.

For the purposes of this project, we define an "outage" as 15% of the meters having voltage less than 18 V at any minute over the course of the year. We set a threshold of 15% because a single meter or even multiple meters droping below 18V hardly signifies a system-wide event, but as that percentage rises the stability of the grid becomes increasingly compromised. Under these constraints our dataset shows an outage 20% of the time, or approximately 1,750 hours out of the year.

We added weather data from forecast.io which we found using the geospatial coordinates in Devergys dataset and focused on cloud cover, which affects solar panel output, and temperature, which influences electrical loads through cooling demands, as our primary weather features.

We used Devergy's data to construct our grid features. To convert the data into features appropriate for machine learning algorithms, we first split it into separate columns for each trait and meter/enbox combination on the grid (ie. Voltage In for Meter 827). We dropped all of the data other than:

- *Vin (mV)*
- *Vout (mV)*
- *I (cA)*
- *Mode (a measure of when the meter has been shut off)*

The cardinality of our feature space, including a month of year feature to account for seasonality, was 808. We interpolated the meter and weather data at one-minute increments, as the minute-level data provides the most granularity about the state of the system. However, when applying machine learning models, we often resampled the data at an hourly level for computational speed.

III. Model Selection & Methodology

By nature, predicting either 'outage' or 'no outage' is a binary classification problem. We therefore focused our analysis on three classification methods: support vector machines, logistic regression, and random forest classification.

A. Logistic Regression

After preparing our featurespace, we chose to begin our analysis by running a logistic regression. Logistic regression was chosen as a starting point for our analysis for several reasons. First, we are attempting to predict the outcome of a binary categorical dependent variable (class labels) based on our set of 808 predictor variables (features). Therefore, we had no reason to believe that our results would not be affine. Second, logistic regression is very interpretable and it is easy to attribute variance to specific features. Given our hypothesis, we presumed that logistic regression would explicitly show the proportion of the variance in our model that was due to weather features.

Logistic regression works by calculating the sigmoid function of a linear combination of the features, tuned with the parameter θ:

$$h_\theta(x) = \frac{1}{1 - e^{-\theta^T x}}$$

The algorithm classifies as example x^i as giving a 1 if $h_\theta(x) > .5$. During the training phase, we choose θ to maximize the likelihood of the training data.

We then decided to run logistic regression again and include weather forecasting at the time of prediction to see if this would better inform our outage predictions.

B. Support Vector Machine (SVM)

SVMs are a type of supervised learning model, and in our case specifically a non-probabilistic binary linear classifier. An SVM functions by mapping examples from different categories and dividing them by a decision boundary separating points by as large of a margin as possible. Mathematically, this can be formulated:

$$\max_{\gamma,w,b} \ \gamma \tag{1}$$

$$\text{s.t.} \ \ y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \ i = 1, ..., m \tag{2}$$

$$\|w\| = 1 \tag{3}$$

By adding an arbitrary scaling constant and using a monotonic transformation on the objective function, we can reformulate the problem as:

$$\min_{\gamma,w,b} \ \frac{1}{2}\|w\|^2 \tag{4}$$

$$\text{s.t.} \ \ y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \ i = 1, ..., m \tag{5}$$

We can solve this optimization using its Lagrangian (more detail in the notes) to find the best possible decision boundary for our data.

In our model we decided to begin testing an SVM classifier with a linear kernel. After training and testing on different datasets for 'Mlandizi', we viewed our results in confusion matrices to asses the accuracy of our predictions.

C. Random Forest Classification

Due to discouraging results from support vector machines and limited success with logistic regression for several hour-ahead forecasts, we decided to explore using a random forest classification to improve and extend the forecast accuracy. Random Forest classification and regression trees (CARTs) are unique among machine learning algorithms because they can handle non-parametric, non-linear, discrete categorical, and continuous data as predictors. Random forests leverage thousands of classification trees at once to improve performance. An additional advantage to Random Forests classifications is that they provide a measure of feature variable importance, an estimate for how important the variable is by looking at how the prediction error changes as the variables changes and all other variables are held fixed.

Random Forests Classification is an ensemble method, where the model is selected as the best among a set (size $\sim$ param n_{tree}) of randomly generated decision trees. Each decision tree is generated on the basis of a different subset of the training data as a randomized, recursive partitioning of the feature set. At each node, the tree algorithm decides which variable among a randomly selected set of variables (size $\sim$ param m_{try}) to split as well as the value of that split. The algorithm decides when to stop at a leaf (as opposed to split again) and which classification to assign to terminal nodes (constant class assigned to each leaf). The out-of-bag dataset (remaining data) is used to estimate the classification error of the model and variable importance of the features. Each vector in the test dataset is classified according to all the decision trees, and the class assignment with the most votes is the classification prediction for the input.

For this application, we used the default values for m_{try} (square root of the number of features, 27) and n_{trees} (10 decision trees). We explored varying the size of the subset of variables from 2 to 45 but were not able to improve the forecast accuracy beyond the statistical variation in the results (prediction accuracy varies about 3% from run to run because of dual randomness inherent in decision tree generation). The number of trees required to bound prediction accuracy and variable importance for a CART model is related to the number of predictors, and we varied n_{trees} from 10 to 200 and found there were significant improvements through 100 trees and then they levelled out (from a prediction accuracy low of 88% with 10 trees to a high of 97% with 100 trees).

As the class frequencies are unbalanced in this classification problem (20% outages, 80% no outages) we explored the effect of overriding the majority rule for Random Forest classification of oob examples with class priors. This modification enabled us to slightly reduce our false negative prediction error but resulted in a non-compensated increase in false positive prediction erorr (from 5% to approxmately 8% for a 1-hr ahead prediction). We elected to proceed with using modified class weights.

IV. Results

To evaluate our models we used **K-fold Cross Validation** with $k = 10$, meaning that we trained on 90% of the model and tested on 10%, calculated the out-of-sample error for 'outage' and 'no outage' predictions, and returned the average of each of those errors across partitions of the data. The 1-hr ahead prediction accuracy for 'outage' and 'no outage' on the grid for each of our models are tabulated in Table I.

Both logistic regression and random forests predicted outages with a less than 10% error for several hours into the future, see Fig. 2.

The accuracy of logistic regression in predicting outages (which is more significant for our partners than predicting non-outages) was fairly reasonable up until the about 4 hours ahead, when error began climbing past 40% and remained noisy up until twelve hours into the future. Because outages are so costly, it is worth the avoided cost for the micro-grid operators to send personnel to check on the state of the system even if the outage prediction is only 40% certain.

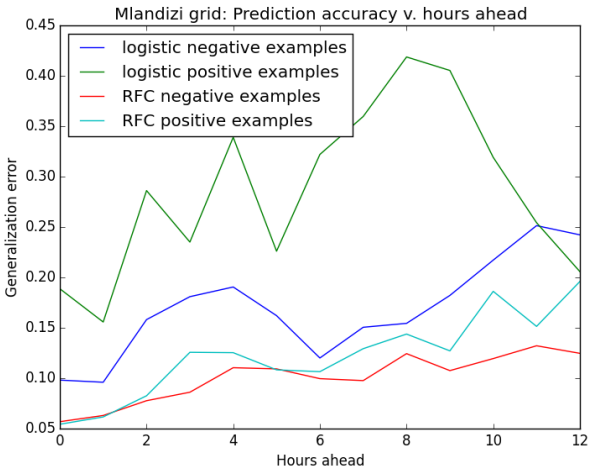

Fig. 2. Outage Prediction Accuracy for Logistic Regression and Random Forest Classification up to 12 Hours in Advance.

However, whereas the logistic regression's outage prediction devolves into near-random noise 3 to 4 hours in the future, random forest classification appears capable of predicting outages up to 12 hours in the future with a generalization error of around 20%. This result is a **significant** improvement over logistic regression, and is far enough in advance to permit Devergy to anticipate and adequately address the low voltage event by either sending personnel to install new generation capacity or limit overall customer power consumption.

It is worth noting how poorly Support Vector Machines performed on our test data in general. First, the number of support vectors often exceeded 75% of the training examples. Seccond, SVMs were unable to capture any of the meaning of our data; the results were so poor we have chosen no to include them in the graph. SVM predicted a non-outage in nearly every case, so its prediction error was at or close to 100% for outages, and 0% for non-outages. We think this reflects the facts that our data is not seperable, so a decision boundary is not appropriate. It's highly probable that a large number of features, rather than a few support vectors, drive the variance together.

TABLE I
Model 1-hr Prediction Accuracy (**K-fold**)

Model	Outages	Not Outages
SVM	100	2.7
Logistic Regression	90.2	81.2
Random Forest ($k = 100$)	93.7	93.9

V. Analysis

Our investigations demonstrated that it was intractable to predict grid outages on the basis of weather alone (temperature and cloud cover) without additional parameters such as the state of charge of the batteries. Results from SVM on the weather-only feature space were not encouraging; the occurence of outages seems nearly independent of temperature and cloud cover. See Fig. 3. As a result, we modified our

strategy and leveraged all features related to the state of the system (current, voltage in, voltage out, and mode) to train our model.

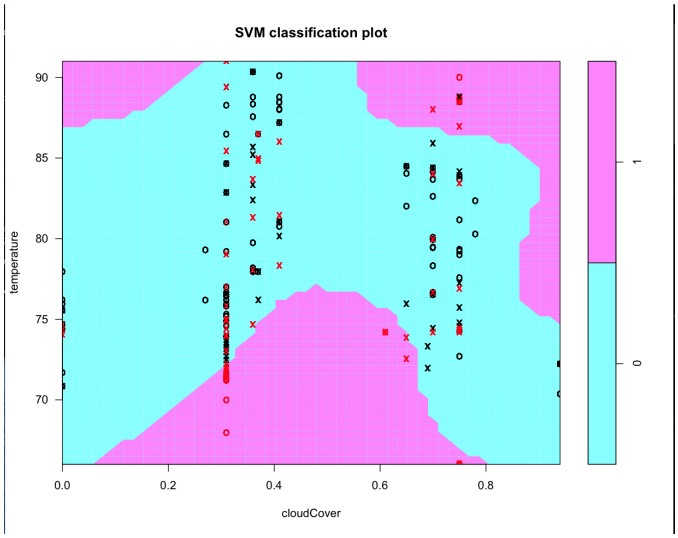

Fig. 3. Poor SVM results on the basis of temperature and cloud cover only.

Using k-folds cross validation to train our model on the full set of 808 features, the prediction accuracy for outages of our logistic regression and random forest models approached 90% for the 1-hour time window. The accuracy of the logistic regression and random forest model precipitously decline at the 3 to 4 hour window (though on different scales). We hypothesize that 3 to 4 hours may be the order of a first-degree lag on the system, i.e. that there are strongly interpretable signals of a near-term low voltage event in the state of the system up to 4 hours in advance, and that there may be a similar jump in the prediction error some additional number of hours into the future.

Based on the results of our random forest classification, we ranked our features according to 'variable importance' and found that the top 50 features all account for a roughly similar proportion of the variance (See Fig. 4), so it appears important that we used a large feature set. Also the most important features are mixed according to type, including current (green), voltage in (pink), and voltage out (light blue) readings of both batteries and meters at distributed locations throughout the microgrid.

This conclusion is buoyed by results from principle component analysis. The variance of the first principal component is 130, the second 120, and the remaining components less than 20. The first and second principal components consist of a nearly even distribution of hundreds of features, the first component consisting of more modes (39 out of the top 50 are modes) and the second component consisting of more voltage readings (31 of the top 50). Both of these results suggest that our feature set is fairly resistance to dimensionality reduction and hence our problem is a good application for machine learning.

To control for the possibility that SVM was ineffective because outages were sparse, we attempted L1 regularization,

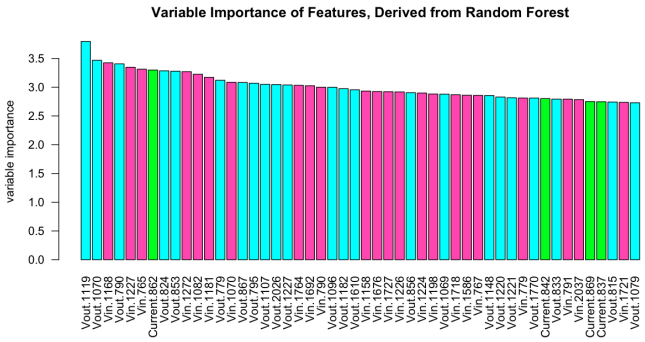

Fig. 4. Ranking of Most Significant Features by Variable Importance.

penalizing a missed outage prediction several orders of magnitude more than a missed non-outage prediction, and training the model on a subset of data where outages were much more prevalent. All of these methods either had no effect or resulted in a model that predicted nearly 100% outages out of sample (as opposed to the ground truth of no more than 20%). As an example of the SVM's bias towards a particular class, see Fig. 5 which plots the false positive and false negative error for training sets with a range of outage percentages.

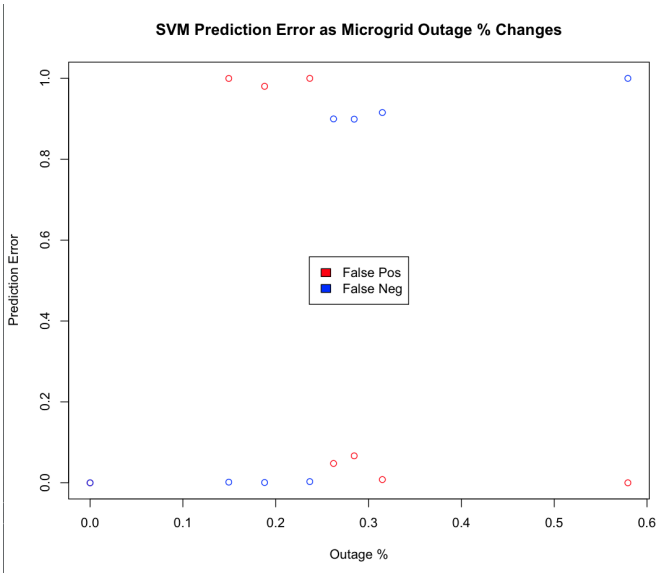

Fig. 5. SVM Models Exhbit Extreme Bias Towards Predominantly Outage or Non-Outage Predictions.

Based on these explorations, and the even distribution of variable importance among features, we suspect that SVM's poor performance compared to the other methods is due to the fact that the variance in our data cannot be explained by a small number of features. We believe this accurately reflects the complex dynamics of the micro-grids. Through our discussions with Devergy, we learned that it is difficult for a human to reliably interpret the implications of a particular meter's measurements in isolation. As an example, a large difference between a node's voltage in and voltage out reading may imply any of: the individual house's load changing,

a box being shut off, a transmission line hiccup, or most critically the start of a low voltage event. These voltage readings are based on a complicated combination of factors that include whether the node is a meter or enbox, which mode the meter is in, and what the potential differences of the nearby nodes are. Without a tractable system of human-intelligble rules to translate meter readings into interpretable events, machine learning applications such as random forest classification that leverage the full richness of the feature set were most successful.

VI. Conclusion

Our analysis has revealed several general lessons about applying machine learning methods to predict micro-grid behavior. First, we were able to predict outages quite accurately several hours into the future, which we think reflects persistence in meter voltages, but had trouble predicting outages beyond that time frame. We believe this is because our data doesnt include the features required to longer-term prediction such as levels of battery storage and usage data for consumer appliances.

Second, no particular feature or small subset accounts for a sizeable portion of the variance in the data, rather in order to predict with accuracy out-of-sample we had to leverage models that incorporate interactions between different features. For this reason, Random Forest Classifications far out-performed the non-ensemble models (SVM, Logistic Regression).

Finally, based on our discussion with Devergy and further research, we expected to be able to predict low voltage events at least partly on the basis of weather parameters. However, we had little success with this method; for all prediction windows (including 12 to 15 hours) temperature and cloud cover did not rank among the 50 most important features. We believe temperature and cloud cover were minimally predictive because the types of electrical appliances in microgrids (lighting, phone charging, TV's and radios) are unlike those in a conventional grid (air conditioning, central heating). Whereas heating and cooling loads are clearly related to temperature, the implications of an unusually hot day on the demand for phone charging is not as clear. Additionally, while cloud cover does certainly reduce the power into the solar panel, the power that the battery provides to the grid is a combination of several factors that we did not include, most important of which is its state of charge. While several days of uninterrupted cloud cover may influence the likelihood of a low voltage event, the cloud cover in a particular minute appears to have little bearing on the system stability.

Next steps include to improve on our feature selection by including autoregressive terms, battery charge levels at supply nodes, and a seasonal parameter (system outages appear to be clustered in the summer). Additionally, we would like to run a more systematic feature selection based on further PCA analysis and our results from the random forests feature ranking.

References

[1] International Energy Agency, *World Energy Outlook 2010*, Paris, OECD/IEA 2010.

[2] Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32.

[3] Liaw, A. Wiener, M. (2002) Classification and Regression by randomForest. *R News*, Vol 2/3, 18-22.