Predicting Power Outages Executive Summary

Julio Caceres, Evan Morris, Aaron Weinberg, Anna Zuckerman

April 20, 2024

Society functions on access to electrical power, and when that access fails, the results can be catastrophic. It's estimated that power outages cost the US economy tens of billions per year [3] and are associated with an increase in overall mortality [1]. They strain the response of emergency services and cause lasting damage to infrastructure. Our project is based on a challenge posed by the ThinkOnward organization with the goal of predicting power outages from extreme rare weather events such as storms. Our goal is to **develop a model which can accurately forecast the intensity of power outages which will be useful to first responders, power companies, individuals, and businesses.**

Two datasets were provided with the challenge. First, we use the EAGLE-I dataset [2] of power outage information consisting of the number of people without power per county in the US between 2014 and 2023, reported every 15 minutes. Second, we use a NOAA dataset [6] of storm events consisting of the start time, end time, location and narrative information of storm events between 2014 and 2024. We supplement these datasets with the ERA5 weather parameter dataset [3, 4], which contains hourly values for a wide range of atmospheric and land-surface parameters (eg. wind speed, ground temperature, etc.). We also aggregate data from the US Census for county shapefiles, the US Energy Information Administration for power grid boundaries, FEMA for county population and area, and the Stanford Data Commons for information about buried power lines. We aggregate all these datasets at the county level and construct timeseries representing each feature and timeseries representing the fraction of people without power in the county. We downsample our timeseries to a 6 hour cadence. This data lends itself to many tasks, but as a simple initial approach we aim to **predict the maximum fraction of people without power at the county level tomorrow**, based on weather data over the past 5 days. We take our test set to be the most recent two years of data (2022-2023). We split the remaining data into a training and validation set using cross validation by iteratively training on an interval of observations and validating on the next observation to avoid data leakage, though the details of this varied somewhat between modeling frameworks.

We experimented with a variety of models, including a linear regression on a vector summary of each timeseries feature, a neural network, and various timeseries analyses. We ultimately decided to compare four models using the framework of the sktime package (a naive model, linear regression, gradient boosted regression, and XGboost) and an LSTM using TensorFlow. We use the mean RMSE across counties in the validation set as our metric for comparison. Despite exploration of different hyperparameters and feature engineering attempts, we were **unable to find a model which performed much better than the naive model**. Our models performed no better than guessing that the target tomorrow should be similar to today. While not inaccurate, this isn't a particularly useful result. This suggests that our modeling framework or our representation of our features may not be appropriate for this modeling task.

Though we did not achieve a very accurate model, we did learn some **key lessons about our data and potential future modeling approaches**. First, we determined that the sparsity in our weather event data, which was a major issue for our models, could be mitigated in future attempts by training on intervals selected to contain significant events. We determined that simplifying the target to predict one value was only useful in allowing us to use a wider range of models, but didn't much improve accuracy. We achieve much better results on certain geographic regions than others, suggesting that further analysis of the shared characteristics of these regions could improve our results. Future work could also benefit from a better treatment of long-term trends like climate change, and a more sophisticated treatment of the spatial correlations across counties. Our explorations will provide a valuable starting point for future modelling approaches towards the important goal of forecasting power outages.

Works Cited

1. Ai, Dan and Crago, Christine and Mullins, Jamie, Heat, Power Outages and Mortality in the United States. April 10, 2025. Available at SSRN: https://ssrn.com/abstract=5212367 or http://dx.doi.org/10.2139/ssrn.5212367
2. Brelsford, C., Tennille, S., Myers, A., Chinthavali, S., Tansakul, V., Denman, M., Coletti, M., Grant, J., Lee, S., Allen, K., Johnson, E., Huihui, J., Hamaker, A., Newby, S., Medlen, K., Maguire, D., Dunivan Stahl, C., Moehl, J., Redmond, D. P., … Bhaduri, B. (2023). The Environment for Analysis of Geo-Located Energy Information's Recorded Electricity Outages 2014-2023 (Version 2). figshare. https://doi.org/10.6084/m9.figshare.24237376.v2
3. Copernicus Climate Change Service (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 (Accessed on 07-MAR-2023)
4. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2018): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 , (Accessed on 07-MAR-2023)
5. LaCommare, K.H., Eto, J.H., Dunn, L.N., Sohn, M.D., 2018. Improving the estimated cost of sustained power interruptions to electricity customers. Energy 153, 1038–1047. doi: 10.1016/j.energy.2018.04.082.
6. National Centers for Environmental Information (NCEI). (2024). Storm Events Database. NOAA National Centers for Environmental Information. Retrieved from https://www.ncdc.noaa.gov/stormevents/