Predicting Power Outages

Project Overview: Predicting Power Outages is based on a challenge posed by the ThinkOnward organization (https://thinkonward.com/app/c/challenges/dynamic-rhythms) 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 power outages which will be useful to first responders, power companies, individuals, and businesses.

Stakeholders: According to ThinkOnward, predicting power outages is of interest to utility companies, emergency responders and hospitals, regulatory bodies, businesses, insurance companies, and the general public. Benefits of predict outages include:

- Enhancing public safety by enabling individuals, hospitals, emergency responders, and vulnerable populations to improve emergency response and prepare for loss of power
- Supporting faster power restoration by allowing utilities to minimize damage to infrastructure and use their repair crews more efficiently
- Reducing economic loss by allowing businesses and people to either implement backup solutions or take steps to preemptively mitigate damages related to power loss
- Preventing grid overload and failures by allowing utilities to preemptively adjust demand-response programs

Description of Datasets: Two datasets were provided with the challenge.

- the EAGLE-I dataset 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.
- the NOAA dataset 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, 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 a few other sources.

- the US Census for county shapefiles

- the US Energy Information Administration for power grid boundaries
- FEMA for county population and area
- 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, and downsample to a 6 hour cadence. Note that datafiles are too large to host on github, but are available on request.

Modeling Task: Our engineered and aggregated 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.

Modeling Approaches: 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 used 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.