Power Outage Prediction 2024 Erdos Project Executive Summary

Team members: Yuba Amoura, Monalisa Dutta, Junaid Hasan, Hatice Mutlu, Nolan Schock
Github: https://github.com/amourayuba/Power-Outage-Prediction

Overview

We use data science to investigate the correlation between extreme weather events and power outages within the United States. Our objective is to develop predictive models to anticipate when these outages might happen. These predictions can help stakeholders to improve outage management and response strategies proactively.

Stakeholders: The stakeholders of our project are power companies, medical facilities, hospitals, insurance companies, and also, research labs can benefit from power outage prediction.

KPI (Key Performance Indicator):

- Precision: number of correctly predicted power outages/total number of predicted outages (True Positives/(True Positives + False Positives))
- Recall: number of correctly predicted power outages/total number of outages (True Positives/(True Positives + False Negatives))
- F1-score = 2*Precision*Recall/(Precision + Recall)

Approach (Data Exploration):

- Our primary resource for weather data is the National Oceanic and Atmospheric Administration (NOAA)'s Severe Weather Data Inventory (SWDI)
- Our primary resource for power outage data is the Department of Energy (DOE)'s data table of reported electric emergency incidents and disturbances.
- We found that lightning, tornadoes and mesocyclones are the three most important severe weather events that yield power cuts, so we trained our models on these three events. Because weather events that do not lead to power outages do not exhibit a wide variation in their properties, undersampling will not lead to a loss of information.
- The first step was to merge the weather and the power outage data. For this, we first used longitude-latitude columns from the weather data and converted them into state and county names using the reverse geocoder package. Then we matched the state and county names and the date columns from both datasets and merged them into a single dataset.
- Next, our dataset was extremely imbalanced, so we used undersampling methods to account for that and prioritized precision/recall/f1-score to measure the quality of our models.
- We worked with the data of the past 9 years from 2015 to 2023.

Models:

To predict the power outage, we built classification models k-nearest neighbours, Logistic Regression, Support Vector Machine, Stochastic Gradient Descent, Random Forest and XGBoost for all three weather events with the data of the past 9 years from 2015 to 2023. Our models used spatial (Longitude, Latitude) and temporal (Month of the year) data as well as a subset of radar predictive features. For the lightning events, we used the number of recorded strikes (and its logarithm) as the feature.

Results & Strategies:

- Tree-based ensemble methods (Random Forest and XGBoost) significantly outperformed all the other ones. That is likely due to the complex decision boundaries of the data, in particular for Month of the Year and Latitude/Longitude.
- The predictive power of the lightning data is significantly lower than Mesocyclones and Tornadoes.
- The data in itself was challenging due to its highly imbalanced nature, and the fact that similar weather conditions in the same place in the same month might often not yield a power outage but occasionally will.
- With the caveat above in mind, the best models using the Mesocyclone and Tornado data can recover 93%/90% of the power outages with a precision of 15%. Alternatively, with a higher threshold that minimizes false positives, they can successfully predict 80% of the power outages with a precision of 37%. These are significantly better than baseline predictions.
- The most important predictive features are by far the Month and Latitude/Longitude.

Future Iterations:

Two straightforward improvements would be to use more years of data and try deep learning for the modelling. These would require more computing resources. Also, one could merge all the three weather datasets and work on the combined effect of different weather events and build the models based on this merged dataset.

While tornadoes, mesocyclones, and lightning proved to be the most important features in these models, further development would ideally incorporate other weather data that are not necessarily extreme, such as wind levels or satellite imagery. We also want to expand the analysis to include the geographical distribution of extreme weather events and power infrastructure. Another nice improvement could be identifying the hotspots of vulnerability and assessing the regional variations in outage risk.

Acknowledgements:

We would like to thank Michael Darcy for their enthusiastic mentorship and helpful feedback during this project. We are also very grateful to Steven Gubkin for the amazing lectures during the boot camp. We also want to thank Roman Holowinsky, Alec Clott, and The Erdős Institute for their support during this boot camp.