HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Predicting Power Outages Utilizing Machine Learning and Sensitivity Analysis

Aadrit Talukdar[1] and Jayson Toweh[#]

[1]BASIS Independent Silicon Valley, USA
[#]Advisor

ABSTRACT

Due to climate change and global warming, power outages have been increasing in frequency and severity. Such power outages not only cause disruptions to day-to-day life but also endanger hospital patients and people using power-reliant medical equipment at home, as well as elderly people relying on cooling and heating. With such an increase in power outages, it becomes imperative to find a solution to mitigate the impacts and frequency of such power outages. This paper analyzes the relationships between power outages and climate factors, such as temperature, precipitation, humidity, wind speed, and solar radiation to determine which are the most significant. Machine learning techniques are used to develop predictive models to identify climate conditions that are likely to cause power grid failures. In addition, this paper also investigates the use of these models in analyzing the resiliency of power grids. This study is focused on the West Coast of the United States. The climate factors and power outage data from 2017-2023 are used in this analysis. Artificial Neural Networks (ANN) were developed to perform the analysis. The results obtained during this research indicated that the ANN models were not effective in predicting power outages based on weather factors for the locations chosen. This is likely due to the highly skewed data sets, as well as existing robustness of these locations against weather related factors. Future work should include applying such modeling on other locations that are more susceptible to weather related power outages as well as investigating other modeling techniques.

Introduction

According to the US Energy Information Administration, customers across the nation experienced power outages averaging 8 hours in length in 2020 (U.S. Energy Information Administration, 2020). Climate change increases the number of severe weather events responsible for power outages. Rising temperatures caused by climate change mean more power is being used by consumers, stressing the grid and increasing the probability of a power outage (EPA, 2023). Between 2000 and 2021, about 83% of reported major outages in the U.S. were attributed to weather-related events, with 58% caused by severe weather, such as high winds, rain and thunderstorms, 22% caused by winter weather, including snow, ice, and freezing rain, and 15% caused by tropical storms and hurricanes ("Surging Weather-related," 2022). The average annual number of weather-related power outages increased by roughly 78% during 2011-2021, compared to the previous decade.

Without power, utilities like lights and refrigerators are not able to function, resulting in food wastage. Without functioning air conditioning or heating, impacts of hot and cold weather become more severe. Other than interruptions in day-to-day life, power outages can shut down power-reliant medical equipment used by hospitals and households, with potentially life-threatening impacts (Suiskind, 2023). Power outages have direct impacts on the United States economy, where weather-related outages are estimated to have cost an inflation-adjusted annual average of $18 billion to $33 billion (Council of Economic Advisors, 2013). Data from various studies lead to cost estimates from storm-related outages to the U.S. economy at between $20 billion and $55 billion annually (Campbell, 2012). Power outages affect businesses in revenue loss, damaged equipment, decreased productivity, ineffective customer care, data loss, and supply chain ripples (Suiskind, 2023).

Because of the severity of these power outages, and their increased frequency due to climate change, it becomes imperative to mitigate their effects. This paper analyzes the relationships between different climate factors and power outages to determine which factors have the predominant impact by modeling their relationship with predictive techniques.

Background

Previous research work has looked into modeling the impact of various factors on power outages. Poulos and Camp (2010) developed a decision support system to identify hazard trees along the transmission line right-of-way (ROW). Wanik et al. (2017) showed the importance of including LiDAR-derived tree height data in outage model predictions during Hurricane Sandy in Connecticut. Kenward and Raja (2014) have done studies that show how power outages have increased over the years due to climate change's impacts. Some research work has focused specifically on hurricane-related power outage modeling (Guikema et al., 2014; Quiring et al., 2014). Shashaani et al. (2018) predicted power outages caused by hurricanes using a multistage classification-regression system. Watson et al. (2022) modeled the impacts of extreme weather events on power grids.

Neural network-based machine learning models have proven to be highly successful in predictions of complex systems. A neural network is a system of interconnected nodes that predicts an outcome based on a series of inputs. The different connections between nodes have weights, dictating how much of an impact that connection has on the final result. In between the input layer and the output layer, a group of hidden layers are added. These models are able to represent the complex relations between multiple input parameters and an output parameter. During the training phase, known input and output combinations are provided to the model. The model compares its computed outputs against the expected output and iteratively updates the weights to minimize the error. After training is completed, the performance of the model is tested with a testing dataset which is different from that used in training.

In this study, we model the relationship between climate factors and the occurrences of power outages using such neural network-based machine learning models. In addition, we will use sensitivity analysis techniques to identify inputs that have the most impact on the outcome of the model.

Methodology

Data Collection

Daily weather data from October 2017 to July 2023 was downloaded from Visual Crossing (VC Corporation, n.d.) for 10 cities on the West Coast: Fresno, CA; San Francisco, CA; San Jose, CA; San Diego, CA; Los Angeles, CA; Tacoma, WA; Bellevue, WA; Seattle, WA; Vancouver, WA; and Eugene, OR. The data included temperature, humidity, wind speed, precipitation, and solar radiation.

For each of the 10 locations, power outages of those cities were downloaded from BlueFire LLC (n.d.). This data included the number of customers, the maximum number of customers who had power outages, the total amount of customer hours tracked, and the total amount of hours customers had power outages.

Data Preprocessing

The power outage data was given as one CSV file, so the data for each city was extracted individually. For the model to be able to classify the power outages, an outage percent was calculated by dividing the total number of hours customers were out of power by the total number of hours tracked. After the outage percent was computed for each day, the datasets for both power outages and weather were combined and aligned by the dates. To measure the impact of the seasons, a new data value was added, grouping every 3 months as a season, starting with December, January and February as Winter. As season data is categorical, a process known as one hot encoding was used on that data. One hot encoding creates multiple individual columns for each category of the categorical data. Then, the category that the data belongs to is assigned a value of 1, and the other columns are set to 0.

Table 1 shows a sample excerpt of the aligned input and output data used in the model. Each row represents the data for a single day. Columns Temp through Season represent the input data of climate factors. Columns Customers Tracked through Hours Tracked Total represent the power outage data. The Outage Value column represents the power outage classification that is used as the output variable for the model. Outage Value is the percentage ratio between Hours Out Total and Hours Tracked Total.

Table 1. Example of Combined Dataset with Weather and Outage Information. Data was collected from 2017 to 2023 from 10 different cities in the West Coast of USA. Season 0 = Winter, Season 1 = Spring, Season 2 = Summer, Season 3 = Fall.

Temp (°F)	Humidity (%)	Precip (in)	Wind Speed (mph)	Solar Radiation (W/m2)	Season	Customers Tracked	Max Customers Out	Hours Out Total	Hours Tracked Total	Outage Value (%)
63.34	48.01	0.000	13.13	195.47	3	5589	2888	8483.21	135043	6.28
65.44	61.40	0.031	13.06	183.03	3	27138	10102	16703.63	638676	2.62
63.53	66.34	0.000	11.57	165.07	3	37597	2526	7319.46	902328	0.81
57.47	65.61	0.018	12.10	126.16	3	38639	7449	10779.74	922656	1.17
59.41	74.77	0.025	11.01	135.30	3	42637	2755	10083.50	1023287	0.99
58.13	73.27	0.111	10.10	110.63	3	42637	1753	7347.55	1023287	0.72
59.83	73.26	0.027	11.36	114.93	3	42637	9560	11968.81	1023287	1.17
53.03	57.34	0.000	11.01	123.46	0	42637	6575	17918.28	1023287	1.75
51.19	55.47	0.000	6.53	122.30	0	42637	2741	10634.02	1023287	1.04

Figure 1 shows two examples of scatterplots between input variables (Temperature and Humidity) and the outage value.

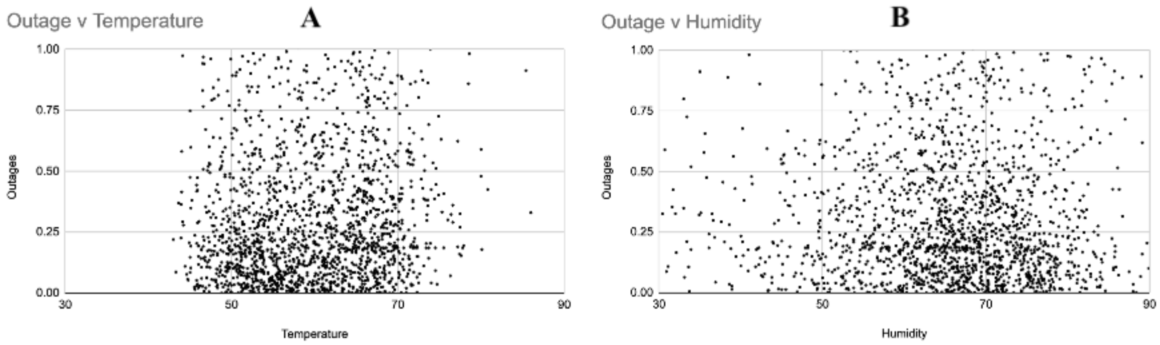

Figure 1. Scatterplots between outage percentage and temperature (A) and between outage percentage and humidity (B)

During the preprocessing stage, different statistics were computed including the average power outage percent, the maximum outage percent, the minimum outage percent, and the percent of the data that had no power outages at all. Along with these statistics, a histogram of the data was made, with the outage percent on the x-axis and the frequency on the y-axis. This was done for each of the 10 locations independently.

Since the model would be affected more by input factors that have large numerical values compared to those with smaller numerical values, it was imperative to normalize all the data to be in the same range. For each of the numerical input parameters, the data was scaled, or normalized, to be a real number between 0 and 1.

Another analysis was done by combining the daily data for every 7 consecutive days to update the scope to a weekly basis. This was done because the daily outage values were very small. In many cases of machine learning models, the model performance was more accurate when such aggregates are used. Such aggregation is applicable in this case as the climate factors are correlated within a few days.

Model Creation

The model used in this project was an artificial neural network for binary classification. It was implemented using the keras.Sequential Python library. As such, the output data needed to be set as either zero or one. To do this, a threshold was set for the outage percent. Anything less than the threshold would be classified as 0 and anything above is 1. In other words, a value of 1 would indicate a day with significant power outages. The combined input and output dataset was split 70-30 among training and testing data. The model was run on each of the 10 locations to study the performance of the model on each location individually.

The model was configured with one hidden layer with ten nodes. For both the input layer and the hidden layer, an activation function of Relu, or Rectified Linear Unit, was used for the nodes (Shaw, 2020). For the final output layer, a Sigmoid function was used. The optimizer of ‘Adam’ and the ‘Binary Crossentropy’ loss function was used for the model’s learning.

Model Performance Measurements

After training and testing the model on the data, the accuracy, precision, and F1-score, which are different metrics to measure the model’s success, were calculated and returned (Bajaj, 2023).

As another metric of success, an ROC curve was plotted using the false positive rate (how often the model classified a non-outage as an outage) and the true positive rate (how often the model classified an outage correctly). Once the curve was plotted, the area under the curve, or the AUC, was calculated. The AUC value represents how well the model was able to separate the two classes (Bajaj, 2023).

Along with these, a confusion matrix was plotted to highlight the classification categories. A confusion matrix gives a representation of the model's successes and errors (Brownlee, 2020).

Sensitivity Analysis

This artificial neural network takes many different input variables. As an experiment to see which climate variables are the most impactful to the power outages, different sensitivity analysis techniques were used. One such method used by Hunter et. al (2000) were adding noise to certain input variables to see the impact on the accuracy. This was achieved using the numpy.random library, creating a Gaussian noise array, with a mean of 0 and a standard deviation of a fraction of the input variable's mean. This noise was added to the input variable, and the model was run with different fractions to see the effect of the noise. This was repeated for each input variable, such as temperature, humidity, wind speed, precipitation, and solar radiation. Because the season data value is categorical, this method of sensitivity analysis was not applied to that variable. Another method of sensitivity analysis used by Hunter et. al (2000) was also used for the sensitivity analysis. This approach replaces all the entries of a specific input variable with the mean of that variable.

For both methods, the accuracy and other metrics were examined. If the performance decreased significantly with the added errors, then that variable on which the sensitivity analysis was used would be important to the model. The larger the reduction, the more important the variable would be.

Results

Data Statistics

The statistics of the computed outage percent data for all 10 locations are shown in Table 2. The average, maximum, and minimum of the computed power outage value were calculated across the entire time span of the data. The Zero Percent column represents how often the location experiences a power outage. A high zero percent would represent that there were some outages on a higher number of days.

Table 2. Power Outage Data Statistics from 2017 to 2023 from 10 cities on the West Coast of USA. The statistics were computed using the outage percentage referenced above.

Location	Average (%)	Maximum (%)	Minimum (%)	Zero Percent (%)
Fresno	0.21	33.27	0.0	1.09
San Francisco	0.53	49.25	0.0	1.42
San Jose	0.97	63.04	0.0	0.38
San Diego	0.66	41.78	0.0	2.41
Los Angeles	0.34	100.0	0.0	82.92
Tacoma	0.13	97.89	0.0	76.78
Bellevue	0.23	32.34	0.0	20.70
Seattle	0.18	100.0	0.0	93.04
Vancouver	0.12	21.59	0.0	28.54
Eugene	0.32	100.0	0.0	9.60

Model Performance

The model performance based on daily data from all 10 locations are shown in Table 3. The Training Accuracy is the accuracy achieved by the model during the Training Phase. The Testing Accuracy, Precision, and Recall were computed during the Testing Phase. The F1 Score is computed from the Precision and Recall. The AUC value is computed from the ROC curves (example: Figure 2).

Table 3. Model Performance Results using daily data samples.

Location	F1 Score	Testing Accuracy	Training Accuracy	AUC	Precision	Recall
San Francisco	0.6218	0.7343	0.7556	0.5371	0.5392	0.7343
Fresno	0.8998	0.9324	0.9311	0.6465	0.8694	0.9324
San Jose	0.588	0.6761	0.6523	0.6322	0.6454	0.6761
Bellevue	0.9077	0.9378	0.929	0.5485	0.8795	0.9378
Eugene	0.8681	0.9107	0.9308	0.6071	0.8293	0.9107
Seattle	0.9703	0.9801	0.9759	0.6413	0.9607	0.9801
Vancouver	0.9625	0.9749	0.9655	0.5919	0.9504	0.9749
Tacoma	0.9878	0.9918	0.9839	0.6959	0.9837	0.9918
San Diego	0.5474	0.673	0.6732	0.5236	0.6981	0.673
Los Angeles	0.9227	0.948	0.95	0.488	0.8988	0.948

The model performance based on combined weekly data from all 10 locations are shown in Table 4, using the same metrics as Table 3.

Table 4. Model Performance Results using weekly data samples. The samples are averaged from 7 days to represent the weekly aggregate.

Location	F1 Score	Testing Accuracy	Training Accuracy	AUC	precision	recall
San Francisco	0.6289	0.6703	0.6698	0.5422	0.619	0.6703
Fresno	0.9022	0.9341	0.9009	0.5647	0.8725	0.9341
San Jose	0.5316	0.5275	0.6604	0.5682	0.5388	0.5275
Bellevue	0.8989	0.9318	0.9163	0.4207	0.8683	0.9318
Eugene	0.767	0.84	0.8793	0.4167	0.7056	0.84
Seattle	0.9657	0.977	0.9403	0.6118	0.9546	0.977
Vancouver	0.9653	0.9767	0.9192	0.6012	0.954	0.9767
Tacoma	0.983	0.9886	0.9754	0.9655	0.9774	0.9886
San Diego	0.5357	0.5385	0.5613	0.4878	0.5335	0.5385
Los Angeles	0.8384	0.8901	0.8815	0.4296	0.7923	0.8901

The ROC graphs were created for all 10 locations. The graphs of Fresno and Los Angeles are shown in Figure 2 and Figure 3 as representative examples.

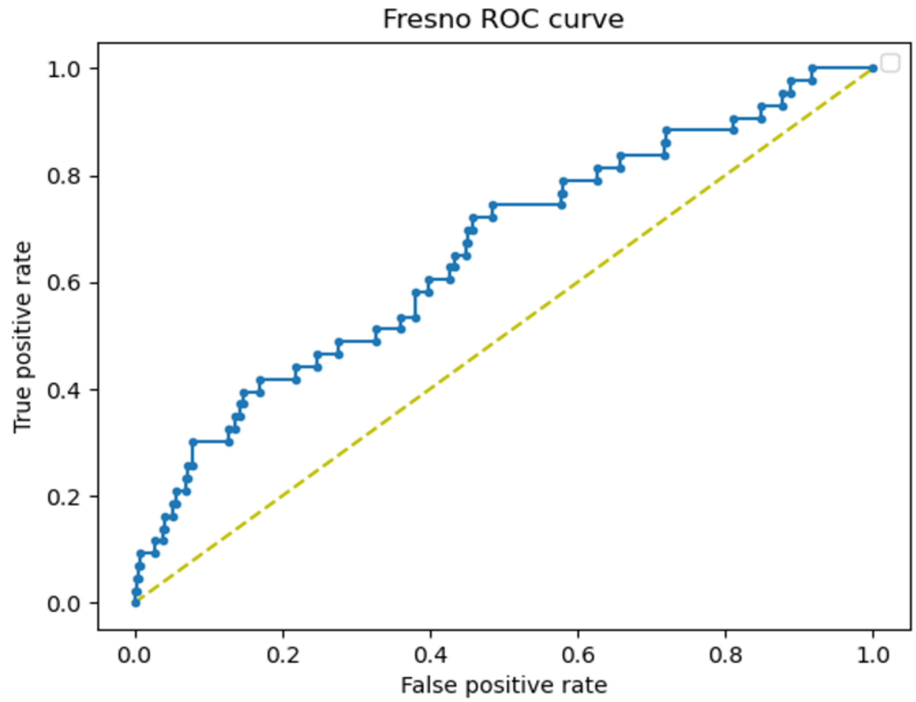

Figure 2. ROC Curve of Fresno from the model trained with daily data samples.

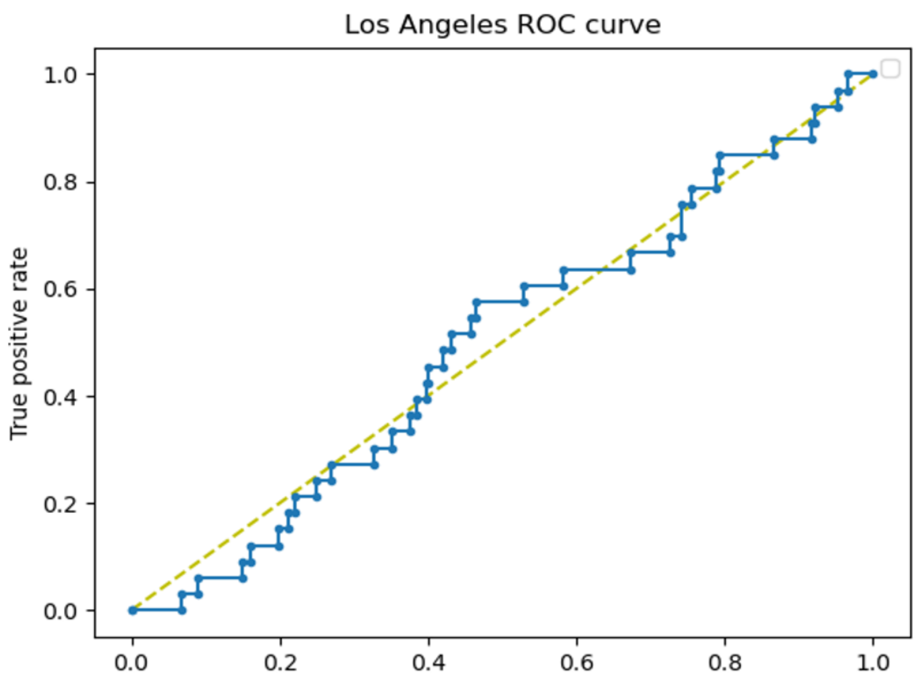

Figure 3. ROC Curve of Los Angeles from the model trained with daily data samples

The Confusion Matrix plots were created for all 10 locations. The plots of Fresno and Los Angeles are shown in Figure 4 and Figure 5 as representative examples.

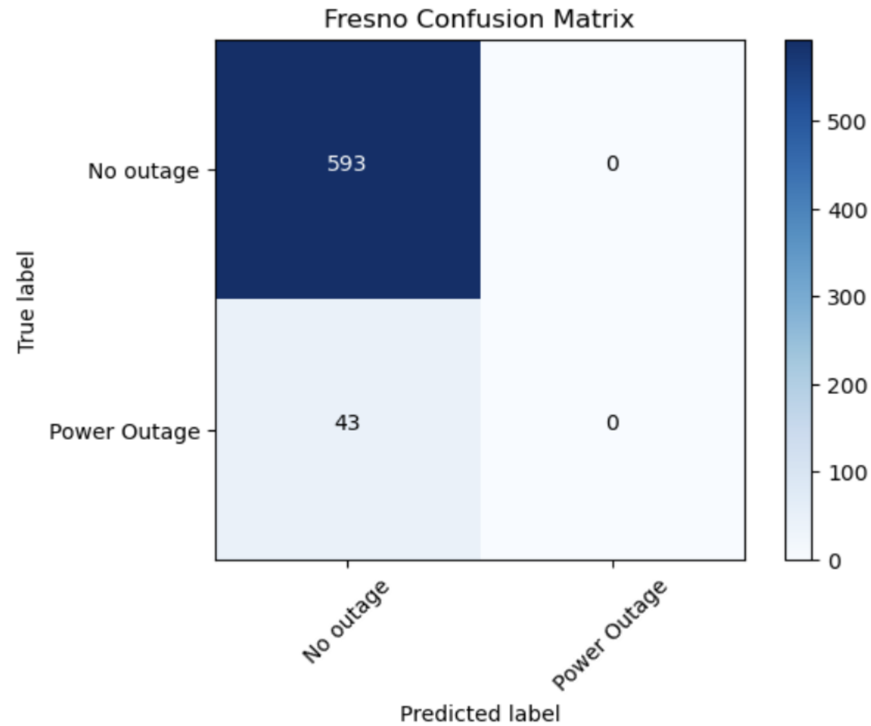

Figure 4. Confusion Matrix of Fresno (Daily)

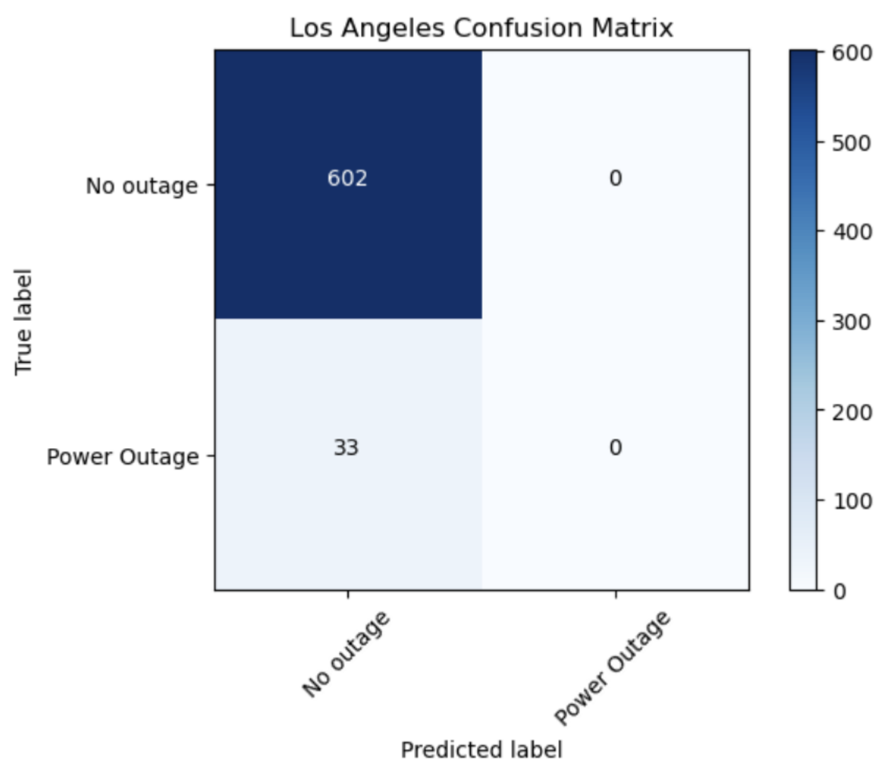

Figure 5. Confusion Matrix Plot for Los Angeles (Daily)

Discussion

Data Statistics

All of the average values for the locations are significantly small, less than 1% (Table 2). This indicates that power outages are usually impacting a relatively small number of customers compared to the location's population. The maximum values indicate that for some locations such as Los Angeles, Eugene, and Seattle, there was at least one day when all the customers experienced a power outage. Locations like Seattle, Los Angeles, and Tacoma experienced any kind of outage for relatively fewer days. Comparatively, San Jose experienced outages for most of the days within the dataset's time frame. But the average and the max are not significantly higher than the other cities. This indicated that even though San Jose experienced outages more frequently, it affected a smaller section of the city. This may be because San Jose is a large city with a wide variation of power grids.

Model Performance

Model performance metrics are shown in Table 3. The primary metric is the accuracy, which is high for most of the locations. But accuracy is not always a good indicator of predictive behavior, especially if the dataset is not evenly distributed. The F1 Score, which represents a combined metric of precision and recall, is high for most of the locations, but the AUC scores were low, closer to 0.5, which indicates high randomness in the classification results (Figure 1, Figure 2). The results from weekly aggregation of the data in Table 4 showed similar results. This indicates that the aggregation did not affect the model performance.

To investigate the low AUC values, a confusion matrix was created and analyzed. The confusion matrices indicate that the model was classifying all test samples as not a power outage (Figure 4, Figure 5). This is because the dataset was highly skewed, with a predominance of no outages. In some areas there were 602 days with no outages, compared to only 33 days with outages (Figure 4). The model ended up classifying all the outage days as no outages. But the relatively fewer number of outage days still resulted in a higher performance metric.

Summary

The results obtained in this study indicate that the artificial neural network model was not able to perform well in the targeted classification task. One likely cause of this is the highly skewed datasets. ANNs usually perform better on well distributed datasets (Brownlee, 2020). Another possibility is that power grids in these locations from the West Coast of the United States were already well protected from weather related elements. It may be that the outages observed in the data set were due to non-weather-related factors, but this needs to be confirmed through further analysis of the data.

The sensitivity analysis was not done as the models did not show good correlations between inputs and output.

Conclusion

In this research project, a deep neural network-based model was developed to predict the possibility of a significant power outage based on climate factors. The performance metrics of the models indicate that the ANN model is not able to accurately predict the occurrence of a power outage based on weather factors in these locations. One primary reason could be that the datasets for these locations were highly skewed. Also, these locations could already have fortified power grids against general climate factors. Future works should include repeating the experiment with different locations from other regions of the United States as well as the world where there is significant impact of non-extreme climate factors on the power grid reliability. The model could be expanded to consider other environmental factors, either natural or man-made, as input variables.

Another future exploration would be to consider alternate machine learning models. Models such as Random Forest and SVMs have been successfully used in other classification problems.

Limitations

The model training was done with data from 10 locations on the West Coast of the USA. The locations were limited due to accessibility of power outage data. Having more locations would allow for a larger sample size, which would benefit the development of the model. Another limitation is that, due to time and computing resources constraints, only an Artificial Neural Network model was used in this study. Other modeling techniques can be used to see if they provide a better performance in predictions.

Acknowledgments

I would like to thank BlueFire LLC for providing the Power Outage Data used in this Project and my mentor, Jayson Toweh for all his guidance.

References

Bajaj, A. (2023). *Performance metrics in machine learning (complete guide)*. neptune.ai. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Brownlee, J. (2020). *A gentle introduction to imbalanced classification*. MachineLearningMastery.com. https://machinelearningmastery.com/what-is-imbalanced-classification/

Brownlee, J. (2020). *What is a confusion matrix in machine learning*. MachineLearningMastery.com. https://machinelearningmastery.com/confusion-matrix-machine-learning/

Campbell, R. J. (2012). *Weather-related power outages and Electric System resiliency*. Congressional Research Service. https://sgp.fas.org/crs/misc/R42696.pdf

Corporation, V. C. (n.d.). *Weather Data & APIGlobal Forecast & History Data*. Weather Data & Weather API | Visual Crossing. https://www.visualcrossing.com/

Environmental Protection Agency. (n.d.). EPA. https://www.epa.gov/climateimpacts/climate-change-impacts-energy

Executive Office of the President. (2013). *ECONOMIC BENEFITS OF INCREASING ELECTRIC GRID RESILIENCE TO WEATHER OUTAGES.* https://www.energy.gov/articles/economic-benefits-increasing-electric-grid-resilience-weather-outages

Guikema, S. D., Nateghi, R., Quiring, S. M., Staid, A., Reilly, A. C., & Gao, M. (2014). Predicting hurricane power outages to support storm response planning. *IEEE Access*, *2*, 1364–1373. https://doi.org/10.1109/access.2014.2365716

Hunter, A., Kennedy, L., Henry, J., & Ferguson, I. (2000). Application of neural networks and sensitivity analysis to improved prediction of Trauma Survival. *Computer Methods and Programs in Biomedicine*, *62*(1), 11–19 https://doi.org/10.1016/s0169-2607(99)00046-2

Kenward, A. (n.d.). *Blackout: Extreme weather, climate change and power outages*. BLACKOUT: EXTREME WEATHER, CLIMATE CHANGE AND POWER OUTAGES. https://assets.climatecentral.org/pdfs/PowerOutages.pdf

Poulos, H. M., & Camp, A. E. (2010). Decision support for mitigating the risk of tree induced transmission line failure in utility rights-of-way. *Environmental Management*, *45*(2), 217–226. https://doi.org/10.1007/s00267-009-9422-5

Quiring, S. M., Schumacher, A. B., & Guikema, S. D. (2014). Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling. *Bulletin of the American Meteorological Society*, *95*(1), 47–58. https://doi.org/10.1175/bams-d-12-00012.1

Shashaani, S., Guikema, S. D., Zhai, C., Pino, J. V., & Quiring, S. M. (2018). Multi-stage prediction for zero-inflated hurricane induced power outages. *IEEE Access*, *6*, 62432–62449. https://doi.org/10.1109/access.2018.2877078

Shaw, S. (2020). *Activation functions compared with experiments*. W&B. https://wandb.ai/shweta/Activation%20Functions/reports/Activation-Functions-Compared-With-Experiments--VmlldzoxMDQwOTQ

Suiskind, B. (2023). *7 important effects of power outages on homes & businesses*. All Dry USA. https://www.alldryus.com/general/effects-of-power-outages-on-homes-businesses/

Surging weather-related power outages. Climate Central. (n.d.). https://www.climatecentral.org/climate-matters/surging-weather-related-power-outages

U.S. electricity customers experienced eight hours of power interruptions in 2020. Homepage - U.S. Energy Information Administration (EIA). (n.d.). https://www.eia.gov/todayinenergy/detail.php?id=50316

United States Power Outage Map. (n.d.). https://poweroutage.us/

Wanik, D. W., Parent, J. R., Anagnostou, E. N., & Hartman, B. M. (2017). Using vegetation management and lidar-derived tree height data to improve outage predictions for electric utilities. *Electric Power Systems Research*, *146*, 236–245. https://doi.org/10.1016/j.epsr.2017.01.039

Watson, P. L., Spaulding, A., Koukoula, M., & Anagnostou, E. (2022). Improved quantitative prediction of power outages caused by extreme weather events. *Weather and Climate Extremes*, *37*, 100487. https://doi.org/10.1016/j.wace.2022.100487