Power Outage Prediction using Weather Data

Abhay Sarda
Boston University
Boston, MA 02134
asarda@bu.edu

Shruthi Sivasubramanian
Boston University
Boston, MA 02134
ssiva@bu.edu

Siva Tejaswi Irkm.
Boston University
Boston, MA 02134
sivairkm@bu.edu

Abstract

Over the last few years, severe weather conditions have caused more than 85% of major power outages in the United States. Events such as hurricanes, tornadoes and hail are frequent contributors of the same. However, calculating the possibility of such an outage would require precise measurements of grid and infrastructure data, which does not exist. By analyzing weather conditions and past historical records of power outages for a single area, it should be possible to abstract away details about the infrastructure and grid. In this paper, we explore such a technique applied to Houston, Texas and Ann Arbor, Michigan, to verify the same. We use the NOAA Daily Summaries and Severe Weather Inventory datasets, coupled with the DOE Electrical Disturbances OE-417 data to predict the possibility of such an outage. To further demonstrate the potential of such a model for transfer learning, we apply it to neighboring states with different infrastructure and weather conditions.

1 Introduction

Power outages are one of the most prevalent issue in developing and developed countries. Finding a way to accurately predict the occurrence of such an event could go a long way in helping people and institutions prepare for the outage and thereby minimize the damage caused due to power failure. For example, diabetes patients who rely on insulin injections need to keep them cooled below a certain temperature. Knowing that a power outage is imminent can allow them to stock up on ice, preventing any medical risk. Hospitals, medical centers and community centers can prepare for severe weather events in a better way if they know a power cut is likely to happen.

Although this analysis would arguably be more valuable in developing countries, which have much more frequent power outages, there is almost no data on power outages or weather data aggregated over a substantial time period for them. Hence, as a proof of concept, we analyzed weather patterns and electric grid disturbances for the United States, specifically.

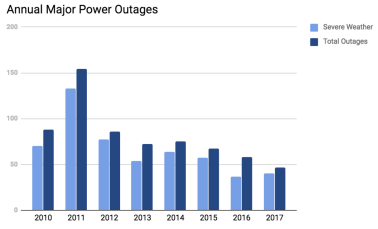


Figure 1: Major power outages between 2010 to 2017. The graph indicates that over 85 percent of the total major power outages in this period occurred due to severe weather events.

1.1 Power outages in the United States

According to the annual electric disturbances summaries from the Department of Energy, an average of above 85 percent of major power outages are caused due to severe weather events (Figure 1).

Out of the several power outages that occurred between the years 2010-2017, Michigan saw 159 power outages and Houston suffered about 100 power outages. Due to the power outage data being so sparse, we picked out cities with frequent fluctuations in weather conditions.

1.2 Approach

Getting data on all these fields and processing them to get meaningful results would have been impossible. Our hypothesis is that for a particular area, such situations would repeat with similar weather conditions, i.e. if an area had a power outage with a particular set of weather conditions, there is a high probability of power outage happening if the similar weather conditions occur again. Doing so allows us to abstract away all the dependencies on the infrastructure, electric suppliers, etc. and should allow us to predict power outages accurately.

2 Background and related works

There has been a lot of work in predicting the cost of a power outage to electric grid providers.[2,3] Some of the factors that have found to affect the probability of such an event are:

- The square of the maximum wind gust speed in a day[1]
- The temperature range in a day
- Precipitation
- Dew Point

Apart from these daily observations, indicators for lightning, thunder and other severe weather events have also been shown to have a dominant effect on the probability[4]. However, most studies have been done for the purpose of analyzing cost to the grid providers, and not to provide a simple metric to consumers about the possibility of power outages.

2.1 Datasets Used

We utilized various weather datasets from Kaggle[7,8,9] and the Climate Data Online Portal, by the National Oceanic and Atmospheric Administration[5]. However, the bulk of our data is from the daily summaries weather data from NOAA. Various parameters such as wind speed, temperature, severe weather indicators and precipitation was recorded daily for stations across the United States.

For the record of power outages, we used the OE-417 Electrical Disturbances records from the Department of Energy[6]. They detail power outages throughout the US going back to 2000. Some of the details mentioned are given below:

- Duration of Power Outage
- States affected
- Number of people affected
- Cause of power outages

For the proof of concept, we decided to analyze the states which have had the highest number of power outages from 2010-2017. The states in descending order of the number of outages are:

- Michigan
- Texas
- North/South Carolina
- California

We chose Houston, Texas as our first candidate for the analysis. Although the daily summaries for all cities have been recorded by centers, very few centers have actually measured all the fields of the observation. For example, the 5 second wind speed which will prove to be crucial for our analysis, was measured by only 2 out of 25 weather centers in Houston. The electric disturbance data was entered into the weather records by creating a new field for it. This gave us a comprehensive data set for our analysis.

3 Data Analysis

The daily summaries dataset has fields for the following parameters:

- Max. 5 second wind speed
- Max. 2 minute wind speed and Avg. wind speed in a day
- Precipitation
- Tmax and Tmin
- Indicators
  - Hail
  - Fog/Ice
  - Heavy/Freezing Fog
  - Thunder

From our literature review, we observed that the wind speed squared has played a dominant effect on power outage probabilities as well. Hence, we added fields for the square of the wind speeds, for all 3 wind speeds. Also, our intuition was to include the delta T parameter in the measurement, since rapid changes of temperature between two days might indicate a severe weather event.

3.1 Dimensionality Reduction

Its imperative to scale the various terms to a consistent range. Doing so would prevent any one term from outweighing all the others. Hence, we scale all the parameters to a range of 0 to 1. After scaling down the data, we need to identify the leading causes of power outages. Reducing the number of dimensions ensures that only the most correlated predictors are analyzed. Doing this involves removing variables which represent redundant data, since the actual contribution of a factor might be divided into several factors, yielding a sub-optimal set of predictors.

3.1.1 Removing multicollinear variables

The need for removing multicollinear variables, stems from the fact that they would convey redundant information and would hence adversely impact accuracy. Hence, we would ideally want to identify variables which represent the same information to a certain threshold, and eliminate those variables. To do so, we calculate the Variance Inflation Factor(VIF) for all the variables, and eliminate ones with VIF above 5.

3.2 Balancing our data set

The very nature of the problem indicates a heavy bias in the number of occurrences of each class. Especially for developed countries, the total number of power outages will be much lower than the number of normal days. For example, in state of Michigan, there were 159 power outages and 2910 normal days in a eight year period, giving a ratio of 1:18. But for solving any classification problem, we need a substantial ratio of both classes.

To solve this problem, we tried out two widely used approaches.

- Undersampling the majority class
  - Reduce the number of 0s in the data
  - Use K-means cluster centroids to preserve key characteristics of the data

  - This would however impact the accuracy of detecting normal days, as there would be many weather conditions that our model has not yet seen

- Oversampling the minority class
  - Increase the number of 1s in the data
  - Synthetic Minority Oversampling technique
  - Using existing points, SMOTE creates new data points through interpolations between existing points
  - Since the new points are created by existing ones, we do not lose any data in doing so.

Oversampling the minority gave the best results, hence we used it for our analysis. Choosing to under-sample the normal days resulted in a decrease in accuracy.

3.3 Spliting the data

The dataset was split into training and testing sets, in the ratio of 2:1. This was done before the dimensionality reduction and balancing the datasets. Doing so prevented any leakage of data between the testing and training dataset. Since we will be relying on oversampling the minority set using SMOTE, we want to prevent the testing set from seeing those data points, as this would give a nearly perfect false accuracy.

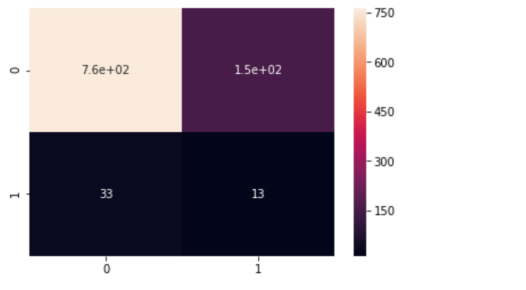
3.4 Defining Performance Metrics

For several reasons, accuracy is not the best metric for measuring our model's performance. For a classification problem with highly unbalanced classes, a model could get extremely high accuracy with just predicting all days as normal days. Another key point to note, is that we want to be able to catch all actual power outages, even at the cost of classifying normal days as power outages. Therefore, we will use the recall score as a performance metric for the same.

4 Model Description

4.1 K-Nearest Neighbors

Since we are aiming to group similar weather conditions together to predict power outages, K nearest Neighbors are a natural choice for the same. Despite their simplicity, the results are quite promising. A large number of normal days were accurately identified, while a lower percentage of power outages were actually identified.



	precision	recall	f1-score	support
Normal Days	0.96	0.83	0.89	915
Power Outages	0.08	0.28	0.12	46

Figure 2: KNN

4.2 Penalized Support Vector Machines

In order to account for the imbalance of the classes, penalized learning can be a valuable tool. Penalized SVM differs from normal SVM by applying a large penalty on classifying actual occurrences as false negatives. However, the results for the same are identical to the nearest neighbors algorithm, i.e. accurately predicting normal days while not predicting power outages.

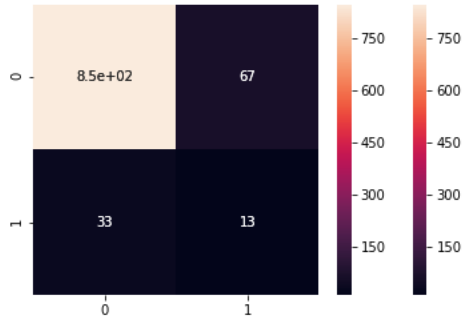


Figure 3: Confusion matrix of Penalized SVM

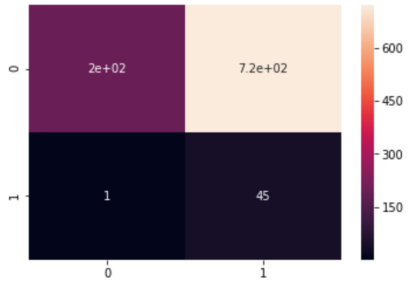
	precision	recall	f1-score	support
Normal Days	0.96	0.93	0.94	915
Power Outages	0.16	0.28	0.21	46

Figure 4: Classification report of Penalized Support Vector Machine

4.3 AdaBoost

AdaBoost offers a markedly different approach than our previous classifiers. It allows for training many weak classifiers and combines their results for greater accuracy. It also allows for different class weights, thereby allowing for class imbalance.

The classifier is able to accurately predict power outages, missing only 1 occurence of it, at the cost of accuracy for predicting normal days.

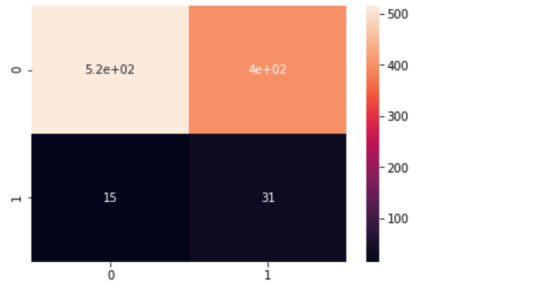


	precision	recall	f1-score	support
Normal Days	0.99	0.21	0.35	915
Power Outages	0.06	0.98	0.11	46

Figure 5: Confusion matrix of Adaboost Classifier

4.4 Gradient Tree Boosting

Gradient Tree Boosting also follows similar design principles to AdaBoost classifiers, relying on random forests of decision trees, but is more robust to outliers due to it's loss functions. Our classifier gave good results for normal days, with a comparable accuracy for power outages as well.



	precision	recall	f1-score	support
Normal Days	0.97	0.56	0.71	915
Power Outages	0.07	0.67	0.13	46

Figure 6: Confusion matrix of Gradient boosting algorithm

4.5 Neural Networks

In order to solve our classification problem, we used one of the most basic non-linear classification model, a simple neural network (Figure 4). Our neural network architecture used one hidden layer composed of 8 nodes, 6 input nodes- for 5 second wind speed squared, fog/ice, 2 minute wind speed squared, avg. wind speed squared, Thunder and heavy/freezing fog- and 2 output nodes which indicate the possibility of a power outage. Our neural network gave us a prediction accuracy of 79% (Figure 3). A plot of the ROC curve and classification report have been shown in Figure 4.

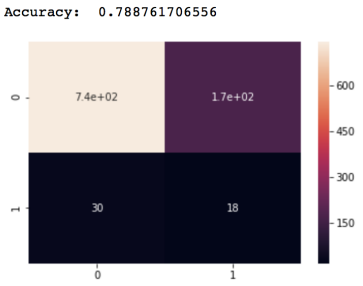


Figure 7: Confusion matrix depicting the classification accuracy of the neural network represented as a heatmap.

5 Transfer Learning

In order to validate the optimal performance of our model, as well as ensuring that we did not overfit our data to our model, we used our model to predict power outages in Indiana, a neighboring state of Michigan.

Using the weather data of a bordering state is one way of validating our hypothesis that power outages are linked to weather conditions. It is highly likely for Indiana to have experienced similar weather events as Michigan given that the states are neighbors, and there is a higher likelihood for the model to accurately predict power outages in Indiana. Therefore, we trained our AdaBoost

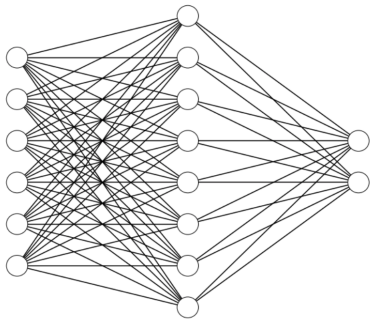
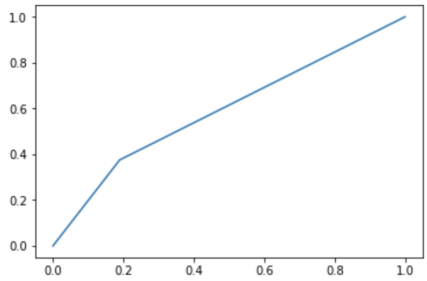


Figure 8: Neural network architecture used for power outage prediction



	precision	recall	f1-score	support
Normal Days	0.96	0.81	0.88	913
Power Outages	0.09	0.38	0.15	48
avg / total	0.92	0.79	0.84	961

Figure 9: ROC curve and classification report

classifier model on Ann Arbor's weather conditions and history, and predicted power outages for Indiana.

Having run our model on on Indiana's weather data, we saw an accuracy of 91%. The confusion matrix, represented as a heatmap has been shown in Figure 10.

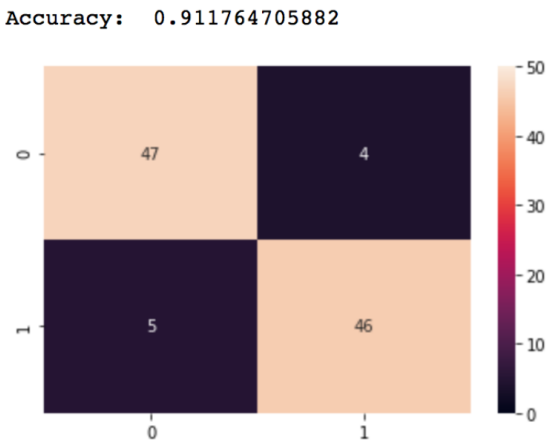


Figure 10: Power prediction in Indiana using model trained on Michigan Data

6 Conclusion

Our model predicts the actual possibility of a power outage quite accurately, at the cost of mislabeling normal days. From our analysis, we can see that our AdaBoost classifier gives extremely high accuracy in predicting power outages, but misclassifies normal days frequently.

- Our model frequently misclassified normal days as power outages
  - For any classification problem, when the minority class is in a much smaller proportion, balancing the datasets can lead to false positives.
  - Which means that even though we are using K means Cluster Centroids, we are still unable to classify normal days with weather conditions that we havent seen before, since its been trained only on 159/2910 normal days.
- The fix for this could be a two layer model, one which accurately catches power outages, with the other using a strong predictor for normal days. Combining the two can give a more balanced approach.

Acknowledgments

We would like to thank Prof. Sang Chin for his support and guidance throughout the project, and Gavin R. Brown for his words of encouragement and advice.

References

[1] Z. Huang, D. Rosowsky, and P. Sparks,*Hurricane simulation techniques for the evaluation of wind-speeds and expected insurance losses,*, J. Wind Eng. Ind. Aerodyn., vol. 89, no. 7, pp. 605617, 2001.

[2] B. J. Cerruti and S. G. Decker, *A statistical forecast model of weather related damage to a major electric utility,* Appl Meteor Clim., no. 51, pp. 191204, 2012.

[3]H. Liu, R. A. Davidson, D. V. Rosowsky, and J. R. Stedinger, *Negative binomial regression of electric power outages in hurricanes,* J. Infrastruct. Syst., vol. 11, no. 4, pp. 258267, 2005.

[4]H. Liu, R. A. Davidson, and T. V. Apanasovich, Statistical forecasting of electric power restoration times in hurricanes and ice storms, Power Syst. IEEE Trans. On, vol. 22, no. 4, pp. 22702279, 2007.

[5]NOAA Climate Data Online: https://www.ncdc.noaa.gov/cdo-web/

[6] Electric Disturbance Events, DOE: https://www.oe.netl.doe.gov/OE417$_annual_summary.aspx$

[7] SPC Reports: https://www.kaggle.com/noaa/noaa-spc

[8] Storm Prediction Reports: https://www.kaggle.com/jtennis/spctornado

[9] Daily Summaries data : https://www.kaggle.com/noaa/gsod

[10] J. Leskovec, A. Rajaraman and J.D. Ullman, Mining of Massive Datasets