Illuminating the Future - Analytical Approaches for Predicting Power Outages in the United States

MGT6203 Team 36 Final Report

July 20th, 2023

Emir Talundzic

Casey Terp

Zoe Yinghong Huang

Gerardo De la O

Ankith Chilkepalli

TABLE OF CONTENTS

LIST OF FIGURES

BACKGROUND

In the United States, the reliance on a stable supply of electricity is paramount for the 331.9 million residents. With the increasing integration of technology into our lives, electricity has become an essential commodity. However, the country has experienced a concerning rise in power outages over the past decade, mainly due to two significant factors: changes in climate patterns and the limitations of the outdated power grid.

The impact of climate change has become evident in the form of more frequent and severe weather events across the nation. Hurricanes, storms, wildfires, and heatwaves have intensified, posing a threat to the power infrastructure. These natural disasters can cause severe damage, leading to disruptions in the supply of electricity and subsequent power outages. Simultaneously, the outdated power grid struggles to handle the growing electrical loads, exacerbating the problem.

Updating the entire national grid simultaneously is not a feasible solution due to its management by over 500 different entities and the significant costs involved. Therefore, grid operators have turned their attention to accurate forecasting and effective load management as critical strategies to ensure proper grid operation, particularly during periods of heavy electricity consumption. Accurate forecasting involves analyzing historical data, weather patterns, and other relevant factors to predict electricity demand accurately. By doing so, grid operators can allocate resources efficiently and maintain a balance between electricity supply and demand. Proper load management techniques, such as load shedding and demand response, play a crucial role during peak usage or extreme weather conditions. Load shedding involves temporarily reducing power supply to specific areas, while demand response programs encourage consumers to voluntarily reduce their energy consumption, reducing strain on the grid.

LITERATURE REVIEW

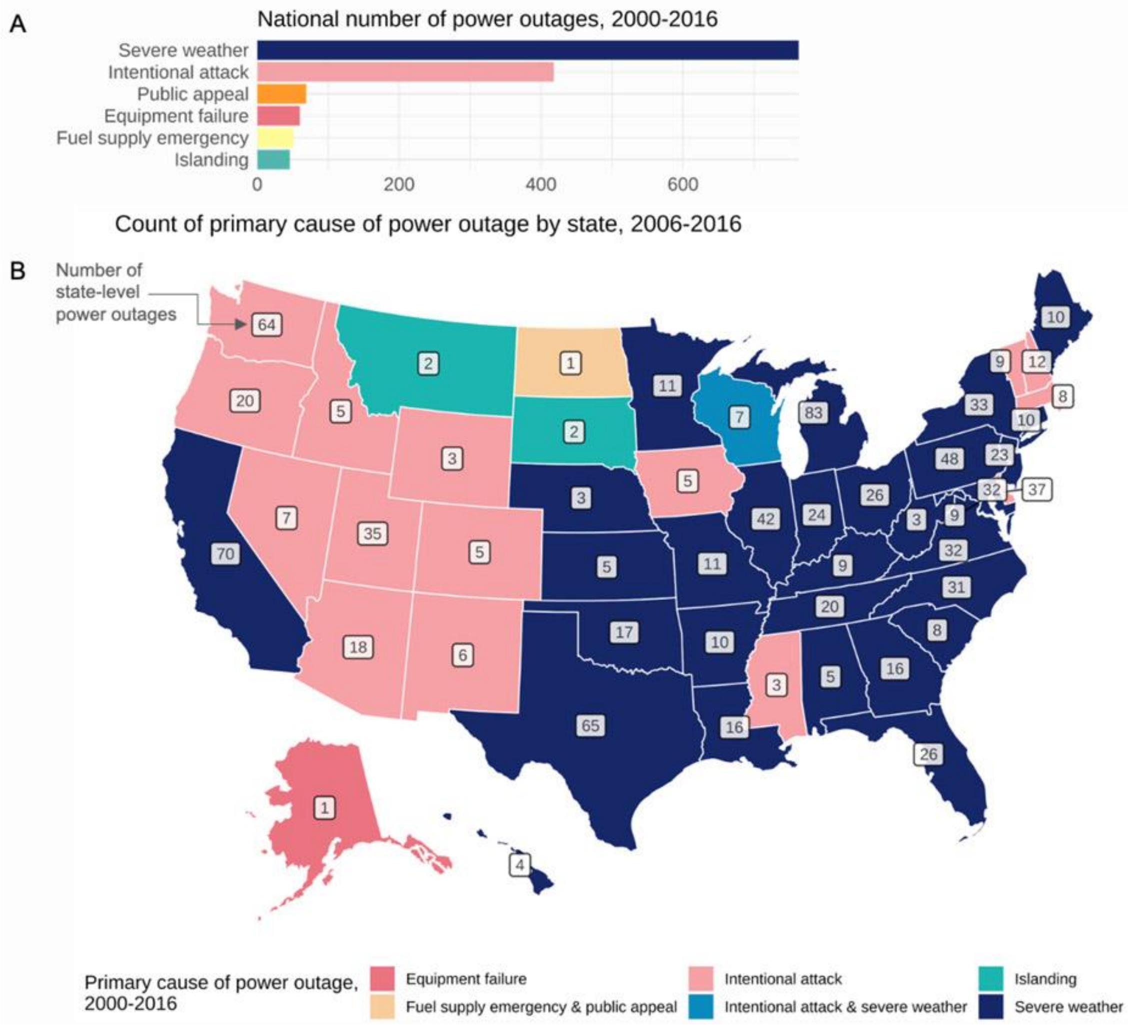
For this project we reviewed several research articles related to blackouts and their significance. We chose to highlight two articles, the first of which is titled *A Survey on Power System Blackout and Cascading Events*[1]. In recent history, several power system blackouts have occurred that left affected populations without power. Outages across the United States and around the world have caused thousands to millions of people to be without power for extended periods of time. Even looking at a small sample of major global blackout events that have occurred shows how the location and scale of these power system failures can vary, and why it’s important to understand blackouts and do as much as possible to prevent them.

These major blackout events can also teach us about some common causes and consequences of blackouts that help inform modeling and analysis to prevent future occurrences. Blackouts can be caused by weather, equipment failure, poor maintenance, human error, protection system malfunction, and more. The consequences of blackouts can be social (affecting medical systems and critical infrastructure), economic (hindering payment systems or transportation ports), or political (impacting national security systems and military operations).

As global energy demand continues to grow, the stability of power systems will be more and more critical to modern life. Power system protection schemes can be implemented to help balance power demand and supply to maintain overall system stability and prevent collapse. Emerging and renewable power generation methods pose challenges for traditional system protection measures, and it is an ongoing effort to maintain power system stability, reliability, and security as the global power system continues to grow and change. It’s important that blackout research and analysis continues to be done to help ensure a reliable power future worldwide.

The second paper we reviewed is titled *Power Outages and Community Health*: *A Narrative Review*. This paper reviewed 50 articles published between 2004 and 2020 to determine the health effects of blackout events on individuals and communities. In the United States, major power outages have increased ten-fold between 1984 and 2012. The frequency and severity of these events will increase with population growth and climate change.

Power outages and blackouts have important health consequences ranging from carbon monoxide poisoning, temperature-related illness, gastrointestinal illness, and mortality to all-cause, cardiovascular, respiratory, and renal disease hospitalizations, especially for individuals relying on electricity-dependent medical equipment. The studies in this field are limited and more work is needed to better define and capture the relevant exposures and outcomes. Future studies should consider modifying factors such as socioeconomic and other vulnerabilities as well as how community resiliency can minimize the adverse impacts of widespread major power outages[2].


A National number of power outages, 2000-2016

B Count of primary cause of power outage by state, 2006-2016

Figure 1: Summary of large power outages in the United States from 2000 to 2016[2]

OBJECTIVE

Problem Statement

We aim to propose a comprehensive set of analytical models that can accurately predict power outages in the grid, benefiting both power utilities and commercial companies. These models enable proactive decision-making, efficient resource allocation, and comprehensive contingency planning. Power utilities can optimize their operations by scheduling maintenance activities and adjusting power generation strategies to minimize disruptions and maintain a stable supply of electricity. Commercial companies can strategically allocate resources, adjust production schedules, and develop contingency plans to minimize financial losses and ensure the continuity of critical business processes.

The implementation of these advanced models empowers stakeholders to be better prepared for power outages. By anticipating and planning for potential outages, power utilities and commercial companies can optimize their operations, reduce downtime, and enhance customer satisfaction. With access to accurate blackout predictions, they can make informed decisions, allocate resources efficiently, and develop comprehensive contingency plans. Ultimately, these models contribute to minimizing the financial impact of power outages and ensuring a reliable supply of electricity for both businesses and consumers.

Approach

To begin the forecasting process for outage events, we undertook a thorough data consolidation process. This involved carefully reviewing and understanding the available data, mapping different data sources, and ensuring the quality of the consolidated data set. During this process, some generalizations and assumptions were made to make our approach feasible during the time allotted for the study. Some of these criteria were:

1. Only CA and NY were considered.
2. Weather data was consolidated from a single weather station and assumed to be state weather.

The next step of the process involved exploratory modeling. Initially, we had planned to use KNN, Exponential Smoothing, and Linear regression to explore the variables and learn as much as possible about load behavior. We soon discovered, however, that there was no load data in support of blackout events, so we were limited to weather and calendar features. The absence of load data reduced our exploratory models to Linear and Logistic regression, which proved to be very valuable in validating our hypotheses.

Since load forecasting was no longer in scope, we implemented a more extensive model selection process to select our final model. Our final approach evaluated Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and Boosted Tree algorithms, coupled with sampling techniques to balance the data set.

Comparing the different models allowed us to identify the most suitable approach, which turned out to be CatBoost model with SMOTE oversampling. Given the nature of gradient boosting algorithms and the effectiveness of SMOTE, this did not come as a surprise. The final model performed significantly better than random in all measures.

INITIAL HYPOTHESES

This study postulated that multivariate models would provide the most insight into grid irregularities. Our underlying hypothesis was based on the premise that unpredictability of input variables introduces inaccuracies in load predictions, which subsequently contribute to blackouts. To substantiate our theory, we put forth the following hypotheses:

- Extreme precipitation events serve as primary indicators of blackouts.
- Deviations in temperature from the normal range within a specific region are indicative of blackouts.
- Calendar variables, such as the month and weekends, are significant indicators of blackouts.

Accurately testing our initial hypotheses presented a major challenge due to the potential mismatch between weather data and blackout regions on a one-to-one basis. To correct this, we assumed that weather data from a specific location could be extrapolated to match the characteristics of an entire region.

Data Processing and Cleaning

The weather data was sourced from the NOAA (National Oceanic and Atmospheric Administration) website. At the start of our project, we were not yet sure of which locations were going to be available from the outage data, so our initial goal was to download daily weather data from all weather stations within the NOAA over the past 10 years. The type of data available was vast, but included features such as daily average temperature, humidity, and precipitation. Daily weather data was available on a per station per year basis in .txt files on the NOAA website that had to be downloaded through an FTP. We used R to access the FTP and loop through all the weather files to download them and merge them into year based .csv files. Once this was done, we downloaded a station key from the NOAA that held location data for all the station IDs. This would allow us to join our weather data to our outage data. While doing initial analysis on the weather data, we noticed that there were not as many stations as we believed there should be. Upon further investigation, the FTP only had a "selected subset" of stations available and did not include enough data from the regions we wanted to focus on. We decided it would be more efficient to manually download the weather data for two locations (New York and California) and use those files for our final weather data set.

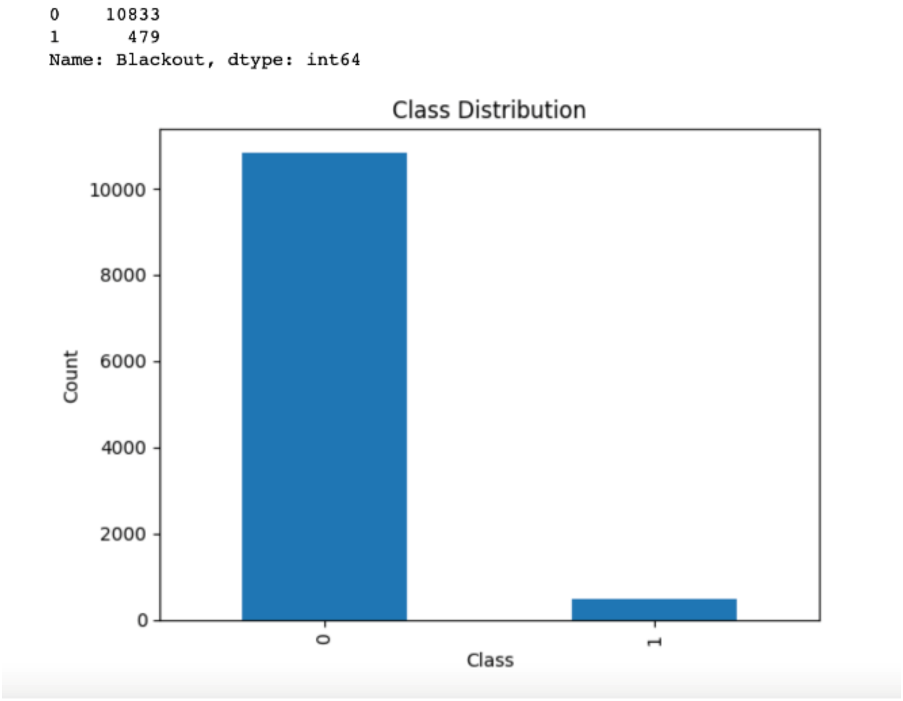
Cleaning of the U.S. outage data was largely handled in python (3.9) relying heavily on the Pandas package (1.5.3). Initial steps were taken to process the timestamp data which corresponded to the total time taken to resolve a blackout event. Timestamp data was parsed to remove text, incorrect entries, and reformatted to a standard date-time (yyyy-mm-dd hh:mm:ss). Further processing was done to separate the time element into four distinct columns: "Date Event Began", "Time Event Began", "Date of Restoration", "Time of Restoration". Next, location data was cleaned once again using Pandas (1.5.3), converting the messy strings of location into two columns: "Affected State" and "Region". The region column was developed from the "Affected State" column, classifying each state into the standard five regions: West, Southwest, Midwest, Southeast, Northeast. Data that was missing or contained incorrect entries was removed from the dataset entirely. Our next steps are to clean the Megawatt and number of customers affected portions of the data to remove comments and other additional text to leave us with cleaned integer data.

Once we obtained the U.S outage data and weather data, the two data sets were further processed to prepare them for merging. Some of the blackout events had multiple entries in the data that corresponded to the same event. These entries were identified and filtered out to ensure that each blackout event had a single entry associated with it. The “Demand Loss” variable contained values for the demand loss that were in multiple formats. We used regex patterns to extract the numerical component corresponding to the demand loss. This was done in a format that could then easily be converted to numeric values. The “Event Type” variable had entries that corresponded to weather with different descriptions and formats. We used several key words to identify the weather-related blackouts and efficiently classify them. We also added the state abbreviations corresponding to each state name. The Weather data contained many columns not properly labeled and instead had codes for the different weather types. The data also had a lot of abbreviations corresponding to different seasons that were unclear. We obtained the corresponding descriptions from the NOAA website and updated the column names and abbreviated data. We then merged the two data sets on the affected state and event date to obtain the combined dataset that was used in our modeling.

As a final pre-processing step, we applied several date transformations to our date column to get additional calendar features. Season, Month, Day of Week, and Weekend were added as features. Then, column names with weather code names were mapped to their appropriate weather condition to make the data set more interpretable. Finally, these features were changed to binary values to facilitate the modeling process. Our final data set consisted of 33 columns and 11,312 rows, with each record representing a full day in each of the two states we selected.

EXPLORATORY DATA ANALYSIS

Before training the model, we plotted the “blackout” column to understand its distribution. Based on the chart below, we inferred that our dataset was very imbalanced.



```
0    10833
1      479
Name: Blackout, dtype: int64
```

Figure 2: Class distribution for blackout

We also plotted the correlation for all the numerical columns within the dataset. There are some columns of weather that are highly correlated, such as Hail and Rain, Mist and Rain, TMAX and TMIN.

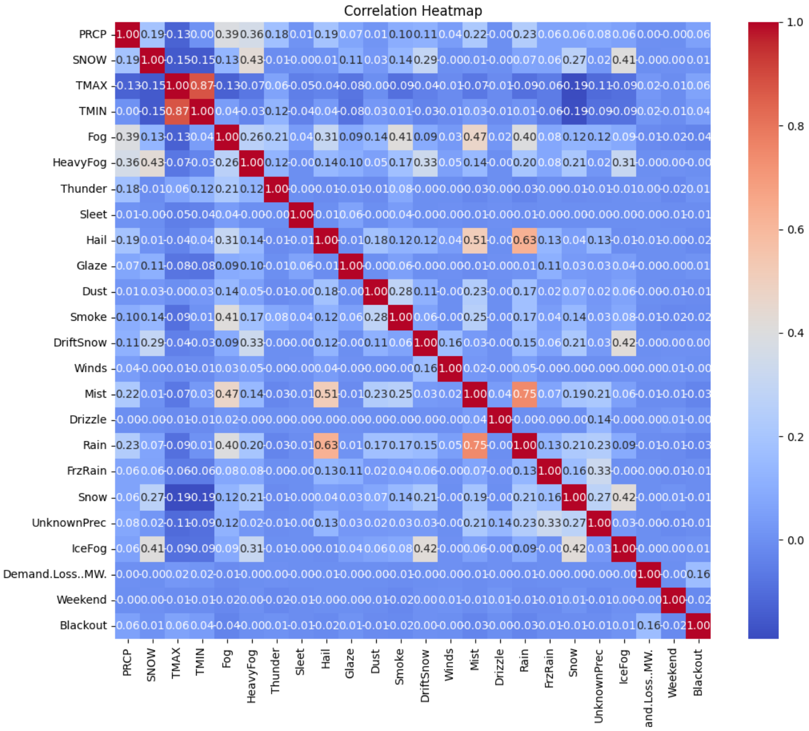
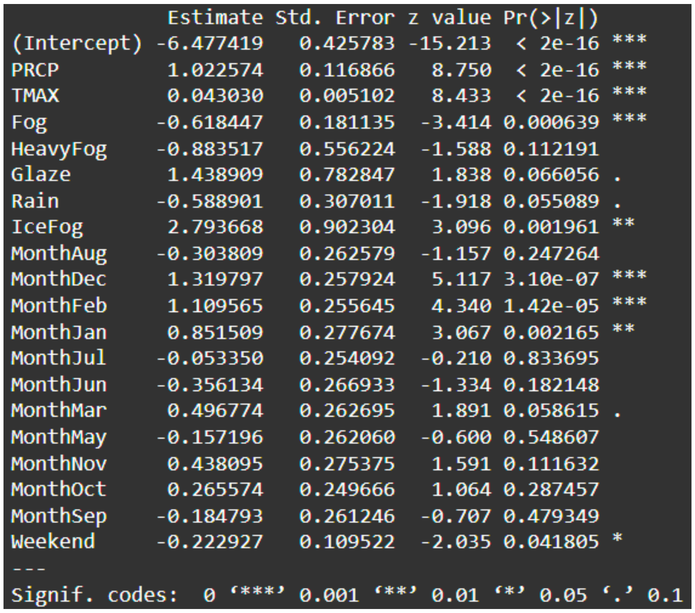


Figure 3: Correlation plot for the predictors

As part of our exploratory analysis, we used logistic regression to test the effect and significance of our predictor variables. We used a step modeling process to find the most important features based on AIC. The results are outlined below. We used these results to validate our initial hypotheses.



```
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.477419   0.425783 -15.213  < 2e-16 ***
PRCP         1.022574   0.116866   8.750  < 2e-16 ***
TMAX         0.043030   0.005102   8.433  < 2e-16 ***
Fog         -0.618447   0.181135  -3.414 0.000639 ***
HeavyFog    -0.883517   0.556224  -1.588 0.112191    
Glaze        1.438909   0.782847   1.838 0.066056 .  
Rain        -0.588901   0.307011  -1.918 0.055089 .  
IceFog       2.793668   0.902304   3.096 0.001961 ** 
MonthAug    -0.303809   0.262579  -1.157 0.247264    
MonthDec     1.319797   0.257924   5.117 3.10e-07 ***
MonthFeb     1.109565   0.255645   4.340 1.42e-05 ***
MonthJan     0.851509   0.277674   3.067 0.002165 ** 
MonthJul    -0.053350   0.254092  -0.210 0.833695    
MonthJun    -0.356134   0.266933  -1.334 0.182148    
MonthMar     0.496774   0.262695   1.891 0.058615 .  
MonthMay    -0.157196   0.262060  -0.600 0.548607    
MonthNov     0.438095   0.275375   1.591 0.111632    
MonthOct     0.265574   0.249666   1.064 0.287457    
MonthSep    -0.184793   0.261246  -0.707 0.479349    
Weekend     -0.222927   0.109522  -2.035 0.041805 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Figure 4: Logistic regression model summary

FINAL MODELS

Data Preprocessing

We started by performing Principal Component Analysis (PCA) on the numeric columns to reduce dimensionality while retaining the majority of variance. We select the principal components that account for at least 95% of the variance and use them as features for training the classifiers.

Test/Train Split

Before training the models, we split the dataset into training and testing sets. We use 80% of the data for training and 20% for testing. The random nature of the split ensures that the models can generalize well to unseen data.

Resampling Techniques

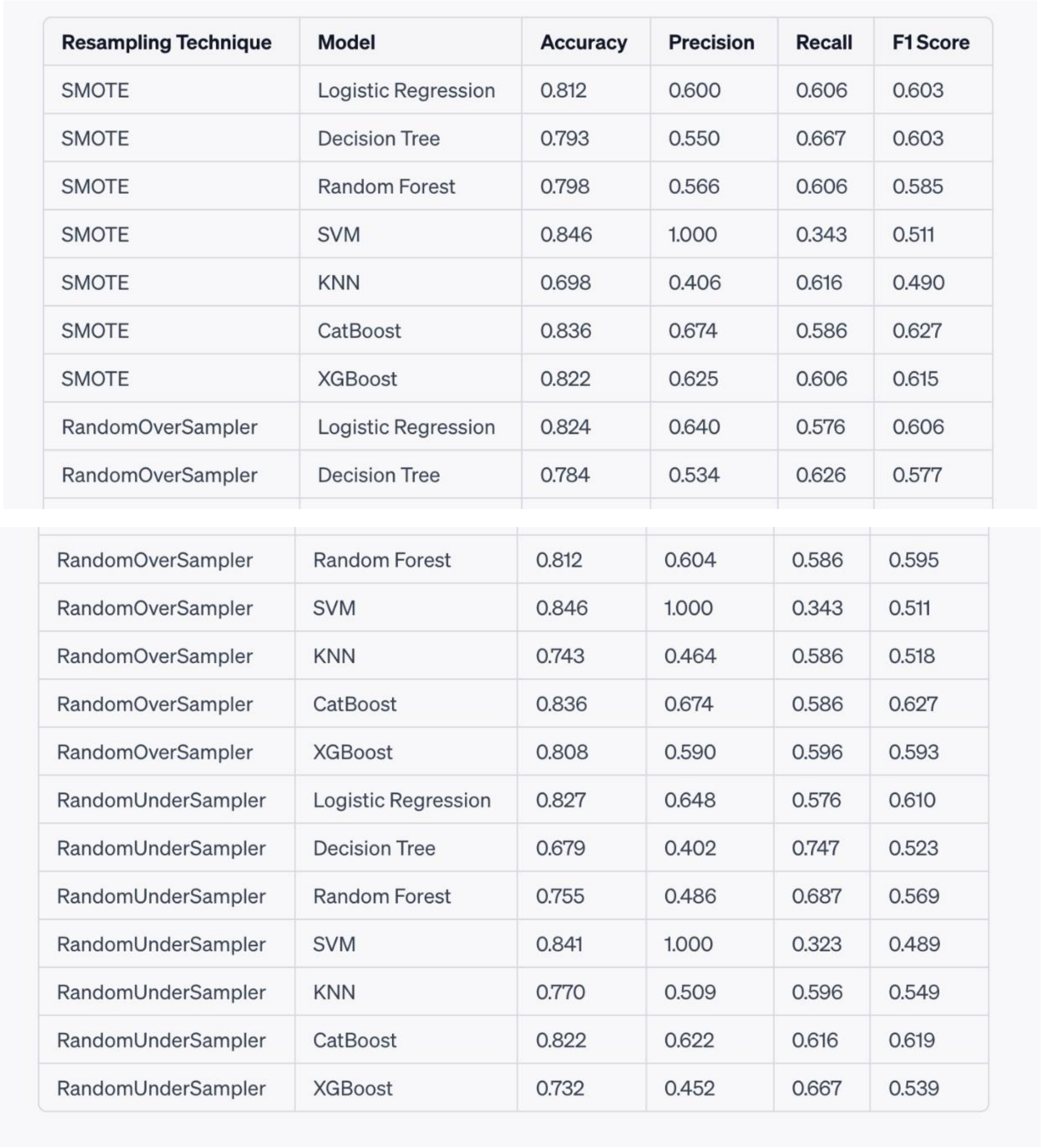
We employed three resampling techniques to address class imbalance in the training data:

1. **SMOTE** (Synthetic Minority Over-sampling Technique): This technique creates synthetic samples of the minority class by interpolating between existing data points. It balances the class distribution.

2. **RandomOverSampler**: This technique randomly duplicates samples from the minority class to balance the class distribution.

3. **RandomUnderSampler**:** This technique randomly removes samples from the majority class to balance the class distribution.

To address the class imbalance issue further, the dataset was divided into two categories: "blackout" (class label 1) and "non-blackout" (class label 0). Additionally, we selected 15% of the "blackout" data randomly selected and combined with all the "non-blackout" data obtain a more balanced dataset and improve the model's effectiveness. Here are the model metrics with this approach:



Resampling Technique	Model	Accuracy	Precision	Recall	F1 Score
SMOTE	Logistic Regression	0.812	0.600	0.606	0.603
SMOTE	Decision Tree	0.793	0.550	0.667	0.603
SMOTE	Random Forest	0.798	0.566	0.606	0.585
SMOTE	SVM	0.846	1.000	0.343	0.511
SMOTE	KNN	0.698	0.406	0.616	0.490
SMOTE	CatBoost	0.836	0.674	0.586	0.627
SMOTE	XGBoost	0.822	0.625	0.606	0.615
RandomOverSampler	Logistic Regression	0.824	0.640	0.576	0.606
RandomOverSampler	Decision Tree	0.784	0.534	0.626	0.577
RandomOverSampler	Random Forest	0.812	0.604	0.586	0.595
RandomOverSampler	SVM	0.846	1.000	0.343	0.511
RandomOverSampler	KNN	0.743	0.464	0.586	0.518
RandomOverSampler	CatBoost	0.836	0.674	0.586	0.627
RandomOverSampler	XGBoost	0.808	0.590	0.596	0.593
RandomUnderSampler	Logistic Regression	0.827	0.648	0.576	0.610
RandomUnderSampler	Decision Tree	0.679	0.402	0.747	0.523
RandomUnderSampler	Random Forest	0.755	0.486	0.687	0.569
RandomUnderSampler	SVM	0.841	1.000	0.323	0.489
RandomUnderSampler	KNN	0.770	0.509	0.596	0.549
RandomUnderSampler	CatBoost	0.822	0.622	0.616	0.619
RandomUnderSampler	XGBoost	0.732	0.452	0.667	0.539

Based on the evaluation of different classification models under different resampling techniques, it appears that the CatBoost model with the resampling technique 'SMOTE' achieved the best overall performance (0.627), with the highest accuracy (0.836), precision (0.674), recall (0.586), and F1(0.627) score on the test set.

Benefits of Class Imbalance Reduction

1. Improved Learning: By reducing the class imbalance, the model becomes better equipped to capture patterns and characteristics present in the minority class (blackout), enhancing its ability to make accurate predictions for blackout occurrences.

2. Mitigation of Bias: Addressing class imbalance helps in mitigating bias towards the majority class and ensures that the model is not disproportionately influenced by the dominant class.
3. Increased Precision: With fewer false positives due to the reduced non-blackout instances, the model's precision in predicting blackout occurrences could be enhanced.

Concerns and Possible Overfitting

1. Overfitting Risk: Overfitting occurs when the model memorizes the training data instead of generalizing well on unseen data, potentially leading to mediocre performance in real-life scenarios.
2. Representativeness: The chosen 15% non-blackout data may not be fully representative of the entire non-blackout class. This could impact the model's ability to generalize to new, unseen data from the non-blackout class.

While our approach of reducing class imbalance by selecting 15% of the non-blackout data and combining it with the blackout class offers initial benefits in improving prediction accuracy, we acknowledge the potential risk of overfitting and the need for a more representative dataset to create a robust and reliable blackout prediction model for real-life business applications. Continued efforts to collect more blackout data and explore various balancing techniques will contribute to the development of a more practical and effective blackout prediction solution.

CONCLUSIONS

Overall Summary

Predicting blackout events using only weather data proved to be a challenging project. Data wrangling for our weather data and outage data was time consuming, and the imbalance in our final dataset gave us many challenges while attempting to create a model that could accurately predict blackouts. Based on the number of models evaluated, we would choose to use our CatBoost model with SMOTE resampling, as it did have the best overall performance. Recall was still relatively low, at .586 so although this model does give us a better chance at predicting an outage event than tossing a coin, it would still need additional work to make it more robust.

Hypothesis Validation

While our final model did not produce the highest accuracy, we were able to validate the following hypotheses:

1. Extreme precipitation events serve as primary indicators of blackouts.
    a. Precipitation and Ice Fog (a form of extreme precipitation) were both statistically significant and increased the probability of a blackout. Ice Fog was also the most predictive variable overall in terms of effect.
2. Deviations in temperature from the normal range within a specific region will be indicative of blackouts.
    a. Maximum Temperature was statistically significant, although its effect was the smallest out of the significant variables.

3. Calendar variables, such as the month and weekends, will exhibit high significance, although their predictive power will be comparatively lower than that of precipitation and temperature.
    a. The winter months were shown to significantly increase the probability of a blackout. Contrary to our original hypothesis, however, the relative effect of December and February turned out to be larger than the effect of precipitation.

This data can prove valuable on its own and can help improve future models for predicting outages.

Future Work

Given more time and resources, there are many aspects of our project we would add or change.

1. Having access to more robust weather data.
    a. Incorporating more samples of weather data from different weather stations in the states chosen might have given us more features to use in our predictions
    b. Having access to weather predictions n days before a date could also have given us more features to use
2. Additional data sources for predictions
    a. While weather data was important, incorporating other data such as daily electrical consumption trends would have made our dataset more robust and could have improved our model accuracy
3. Including Brown Out data
    a. While we focused specifically on blackouts in this report, Brown Outs are also a severe problem. Being able to include brownouts could have also made out dataset more balanced.

DATASETS

To understand the structure of the data provided, please refer to the readme file in GitHub. In the instance that they cannot be downloaded within the zip, please download them from the following link:

https://drive.google.com/drive/folders/1DqBCvKOm5B-JDlRBch91zshn_fDEuVcQ?usp=sharing

CITATIONS

1. Alhelou H, Hamedani-Golshan M, Njenda T, Siano P. A survey on Power System Blackout and Cascading Events: Research Motivation and Challenges. MDPI. https://www.mdpi.com/1996-1073/12/4/682. Published February 20, 2019.
2. Casey J, Fukurai M, Hernandez D, Balsari S, Kiang M. Power outages and community health: a narrative review, PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7749027/. Published Nov 11, 2020.