SMU Data Science Review

Volume 6
Number 1 *Spring 2022*

Article 5

2022

Analysis of the electric power outage data and prediction of electric power outage for major metropolitan areas in Texas using Machine Learning and Time Series Methods

Renfeng Wang
Southern Methodist University, wangrenfeng0@gmail.com

Venkata Leela 'MG' Vanga
Southern Methodist University, vvlmg332@gmail.com

Zachary B. Zaiken
Southern Methodist University, zzaiken@gmail.com

Jonathan Bennett
Southern Methodist University, jonathan@konoanalytics.com

Analysis of the electric power outage data and prediction of electric power outage for major metropolitan areas in Texas using Machine Learning and Time Series Methods.

Zachary Zaiken[1]
zzaiken@gmail.com

Renfeng Wang[1]
wangrenfeng0@gmail.com

Venkata Leela Maruthi Ganesh Vanga[1]
vvlmg332@gmail.com

Jonathan Bennett[2]
jonathan@konoanalytics.com

[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2] KONO ANALYTICS, 1510 Tulane St, Houston, TX 77008 USA

Abstract. – With growing energy usage, power outages affect millions of households. This case study focuses on gathering power outage historical data, modifying the data to attach weather attributes, and gathering ERCOT energy market conditions for Dallas-Fort Worth and Houston metropolitan areas of Texas. The transformed data is then analyzed using machine learning algorithms including, but not limited to, Regression, Random Forests and XGBoost to consider current weather and ERCOT features and predict power outage percentage for locations. The transformed data is also trained using time series models and serially correlated models including Autoregression and Vector Autoregression. This study also focuses on traditional machine learning models that assume sample independence when compared to those that assume serial correlation. The results show machine learning models that utilize both weather features and ERCOT data yield a lower RMSE and higher prediction accuracy than using one feature-set exclusively. In addition, multivariate Vector Autoregressive models have lower RMSE compared to univariate Auto-Regressive, univariate Random Forest and univariate neural network models when weather and ERCOT data are included to predict power outages. Top performing traditional machine learning models are packaged into an external facing web application for public use in determining current power outage risk.

1. Introduction

In today's energy dependent world, electrical power outages or interruptions can have catastrophic consequences [1]. Electrical power grid outages have been a topic of research for decades in both physics and engineering [2 – 6]. The North American

Electric Reliability Corporation (NERC) is responsible for effective and efficient bulk power supply for the North American Continent, United States, Canada, and Mexico [7]. Based on the Distributed Energy Resources Task Force Report in February 2017, the entire bulk power supply grid system of NERC is divided based on the following regions: i) Western Electric Systems Coordinating Council (WECC); ii) Midwest Reliability Organization (MRO); iii) Southwest Power Pool Regional Entity (SPP RE); iv) Texas Reliability Entity (Texas RE); v) Northeast Power Coordinating Council (NPCC); vi) Reliability First (RF); vii) Southeast Reliability Corporation (SERC); viii) Florida Reliability Coordinating Council (FRCC). The bulk power supply regional entities are depicted in Fig. 1.

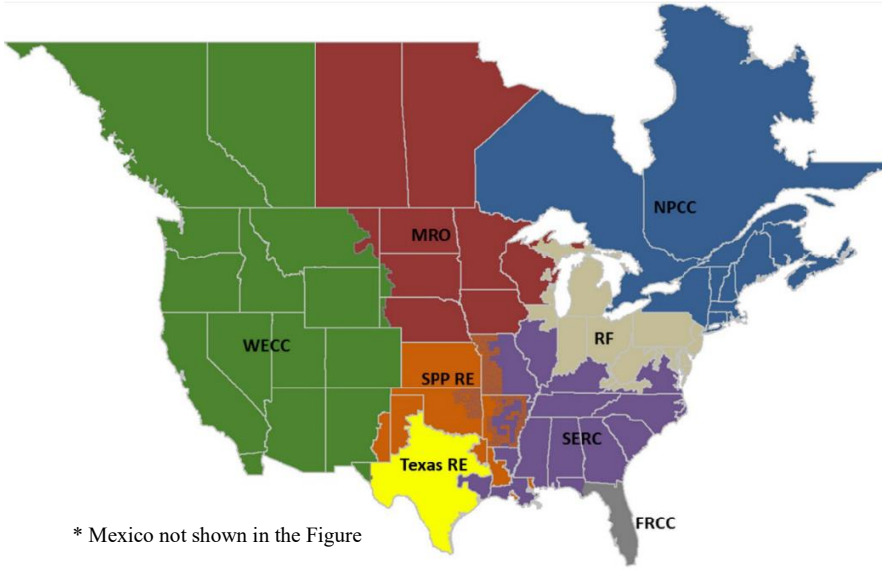


Fig. 1. NERC Bulk Power Supply with eight Regional Entity Boundaries [7].

Electrical power outages are analyzed from a data science and machine learning perspective, focusing on the Dallas-Fort Worth (DFW) and Houston metroplexes in Texas. The Electric Reliability Council of Texas (ERCOT) [8] is the major Texas Regional Entity in charge of autonomous electrical grid system. A primary focus is using data provided by ERCOT.

ERCOT is tasked with maintaining and ensuring system reliability for 90% of the Texas electric load [8]. With over twenty-six million customers and 46,000 miles of transmission lines, it is one of the largest and oldest Independent System Operators (ISO) in the United States. As a deregulated energy market, it provides publicly available financial settlements and market conditions reporting for the Texas competitive wholesale power market [9]. DFW and Houston, two of the largest and fastest growing metro areas in America, make up 47% of Texas population as of 2019 [10-12].

This research also focuses on weather patterns. Over recent years the number of power outages and brownouts across Texas have been on the rise. Extreme weather

conditions—defined as once-in-a-lifetime events—such as: excessive heat, hurricanes and winter freezes put strain on the electricity grid and are occurring more regularly [13]. At 11:00 PM EDT August 25, 2017, Texas was hit by category 4 storm "Hurricane Harvey". The following morning there were 258,137 customers across Texas experiencing power outages [31]. It is the costliest tropical cyclone on record, inflicting $125 billion of damages and claiming over 100 lives. The resulting flooding affected hundreds of thousands of homes that displaced 30,000 people and required 17,000 rescues [32]. The storm also resulted in financial consequences for the energy industry with temporary closure of onshore/offshore oil production, petroleum refineries, natural gas processing plants and ports [31]. During the 2021 winter storm Uri, temperatures plunged below freezing levels in Texas from February 14th-20th. More than two thirds of Texans lost power, 49% experienced disruptions in water service and a reported death toll of 210 people. Additionally, the state experienced financial losses estimated between $80 billion to $130 billion [33].

In addition, complex interactions on the wholesale market can create electricity supply and demand problems that lead to unexpected outages [14]. The producers of power in ERCOT earn revenue primarily from the sales of energy services. They decide whether to keep, retire, or build new power plants based on the investment and ongoing cost relative to the prevailing energy prices and forward-looking market [34]. As such, high prices during times of power scarcity of high demand are a crucial feature to how ERCOT operates. The market design incentivizes long term investment for power supply [34]. Although, this can lead to power supply issues as generation owners are quick to retire or turn off power generation assets when they become uncompetitive or unprofitable [34]. Furthermore, unlike other regions in the U.S., ERCOT does not require mandatory reserve capacity for power. Reserve generation is based solely off generators' decisions to run and customers' decisions of how much and when to consume energy. ERCOT wholesale prices are driven by this supply and demand – where high prices indicate scarcity and low prices indicate energy to the grid is oversupplied [34]. The power consumers in selected Texas counties, though aware of the risks of power outages, are for the most part unaware of when such outages are at highest risk to occur. Therefore, this publication uses data science methods to aid consumers with a power outage prediction tool.

A web application tool is created that predicts the real-time power outage risk percentage for Dallas-Fort Worth and Houston metro areas (including surrounding counties). Furthermore, the web application includes a user interface combining current weather and market conditions to help consumers to predict real time outage risk percentage. In addition, focus is also on building models assuming sample independence (machine learning models) and serial correlation (time series models). Univariate models are compared to multi-variate time series models showing that the inclusion of weather data and ERCOT market predictions increase power outage prediction accuracy. The study also shows how non-linear time series models such as Random Forest and Neural Networks can be leveraged to increase accuracy compared to linearly based Autoregressive models. Traditional machine learning models including Regression, Random Forests, XGBoost, Neural Networks and SVM are trained and tested to predict the power outage percentage for each of the selected Texas counties using weather data and ERCOT market conditions. Additionally, models based on serial correlation including univariate Autoregression, univariate Random

Forest, univariate Neural Networks and multivariate Vector Autoregression are used to predict power outage percentage.

The power outage data is collected using a combination of i) influential results from ERCOT [8], ii) power outage data from 'Poweroutage.us' [26], and iii) the weather patterns from 'Openweathermap' [27]. The collected data have been merged according to real-time stamp and are stored on a relational database (Microsoft SQL Server). The data are further imputed and used to train the traditional and deep machine learning models. A brief description of the data is provided in Appendix - Table 1. The response variable in the dataset—power outage percentage—is highly sparse and imbalanced, with only 1.2% of records with customers experiencing power outages greater than or equal 1% of county population. Techniques including weight scaling, hyper-parameter tuning and transformation of the response variables are explored to ensure consistent model accuracy. The data is transformed to a usable format and used to train the traditional machine learning models and time series analysis. The results of the research show that models that including both weather features and ERCOT market data have higher prediction accuracy and yield a lower RMSE than using only one of the feature-sets exclusively. After using regression algorithms, power outage percentage is assigned into four different classes for classification prediction. Precision Score and Recall are used for performance evaluations. In addition, serially correlated models have lower RMSE with inclusion of both feature-sets than compared to univariate modeling. Weather conditions and ERCOT market conditions are leveraged to yield higher prediction accuracy when determining power outage risk in DFW and Houston metropolitan areas.

2. Literature Review

This research aims to address the problem: what is the current likelihood or risk of DFW and Houston area residents to experience a power outage? Moreover, the focus is to confirm what factors and datapoints are useful or influential in determining when the risk of power outage is high. Likewise, the focus is on what methods and prediction models achieve the most accurate results in assessing this risk. In addition, the paper builds upon the work and findings of various publications summarized below. These studies review current knowledge of the problem, propose prediction methods and selected features to explain the power outage models.

Andersson G et al. in reference [1] studied the factors that were the cause of cascading outages of transmission and generation facilities in the North American Eastern connection, Denmark, Southern Sweden, Scandinavia, Italy, and Central Europe in 2003. The authors used methods like the historical data analysis, deterministic simulation, probabilistic simulation, and high-level statistical models to determine the root causes for grid blackouts. Additionally, they proposed remedial methods that can be implemented. The current research effectively utilizes the insights provided by the authors in reference [1] to implement the methods for the DFW and Houston power outage data.

Carreras B. A. et al. in reference [2] analyzed various studies from 1984 to 2006 using the statistical variations of blackout size, time correlations, and waiting times of

returning power to the Eastern and Western interconnections of the North American (NERC) grid. The authors also studied risks of blackout sizes, as well as the randomness of blackout initialization events and other factors that influence the power outages. Their statistical quantification and time correlations are analytically considered in the present study.

In reference [3] Witthaut D., and Timme M., studied robust synchronization of the grid and the effects of adding new links to the grid. The authors determined that adding the new links were counter-intuitive and compared the phenomena to Braess's Paradox in traffic networks. These insights were considered while collecting the data for power outages.

Pahwa S. et al. in reference [4] studied the renewable energy distribution and the development method for load growth and power fluctuations in the network. The authors also considered blackouts with increasing network size and recommended remedies for the power failures. The recommendations from the authors are considered to build the power outage predictions.

Vaiman M. et al. in reference [5] studied the causes of power outages and the discussed the engineering factor that affect the power outages. The following are the prominent engineering factors that are considered for this case study: i) overloaded transmission lines that subsequently contact vegetation; ii) overcurrent/undervoltage conditions triggering distance relay actions; iii) hidden failures or inappropriate settings in protection devices, which are exposed by a change in operating state; iv) voltage collapse; v) insufficient reactive power resources; vi) stalled motors triggered by low voltages or off-nominal frequency; vii) generator rotor dynamic instability; viii) small signal instabilities; ix) over (or under) excitation in generators; x) over (or under) speed in generators; xi) operator or maintenance personnel error; xii) computer or software errors and failures; xiii) errors in operational procedures. These engineering factors are considered as part of domain knowledge to predict power outages.

Ji, C. et al. in reference [6] studied the weather factors and determined that extreme weather is not the only cause for power outages, but rather amplifies existing vulnerabilities that are cloaked in daily operations. The authors claim that lack of failure detail and recovery data has hindered the studies. This study considers the aforementioned observations and use the comprehensively collected power outage data for the analysis.

Biswas, S. and Goehring, L in reference [15] studied the data science models for outage data between 2002 – 2017. The models are used to indicate proximity to failure points and forecast probabilities of major blackouts with a non-intrusive measurement of intermittent grid outages. The approach by the authors to predict the proximity points and probabilities of major blackouts are considered in modeling and feature creation.

Haifeng S. et al. in reference [16] studied the power outage correlation studies from Twitter data including the load and power outage dependence. The current research uses the methodologies implemented by the authors for performing the correlation studies.

Carlsson, F., and Martinsson, P. in reference [17] performed a selected experiment using a random parameter logit model on the willingness to pay high electricity price to avoid power outages. The experiment was conducted based on Swedish electricity market data. The current study considers the approach to use the random parameter logit model to predict the power outages.

Carnero, M.C., and Gomez, A. in reference [18] studied the electric power distribution outages for both health care and non-health care industries using the multicriteria Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH) and Markov chains. The results for the power distribution were compared to the operating maintenance policies in the organizations and alternative plans produced from the machine learning models have been obtained. Both these techniques can be used in classification of the power outages and are implemented in the current study.

Flamenbaum R. D. et al. in reference [19] studied the equipment failure predictions using Random Forest approach for Southern California based on historical data and the incorporation of environmental and geospatial factors. The results emphasize the high possibility of a predictive model while discussing the limitations. The Random Forest approach can be tailored to fit the current study in creating predictive models.

Wang, Deng, C., and Wang, S. in reference [20] studied XGBoost classification which provides parallel tree boost and is illustrated with different classification dataset examples. This article discusses how the algebraic derivation and first/second-order derivatives of the loss functions contributed in XGBoost algorithms. The classification results are measured by the Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves. By examining the results of XGBoost, the package has great potential to apply to regression, binary, and multi-class classification problems when research data are in large scale and their classification labels are imbalanced. The XGboost package is implemented in the current study to handle imbalance datasets without using under sampling or over sampling methods.

In the Warsono et al. publication of the International Journal of Energy Economics and Policy, Vector Autoregressive models (VAR) with specific focus on the uses of multiple variables, and endogenous and exogenous variables are studied [21]. It is determined that VAR can explain the relationship between variables, the impact of one variable on a set of others, and can predict and forecast time series data. VAR is successfully used to forecast the closing prices of energy stocks. The present research outlines how multivariate time series modeling (such as VAR) is more powerful and can produce higher accuracy than univariate time series alone [21]. The current research uses the insights from reference and VAR models in power outage predictions.

Additional methods provided by Kane et al. in the BMC Bioinformatics Journal review the comparison of Random Forest time series models and Auto Regressive Integrated Moving Average (ARIMA) models to predict Avian Influenza H5N1 outbreaks are studied [22]. Some of the flaws of traditional ARIMA models when applied to real world problems, such as linear relationships between variables and assumptions of stationarity, are noted. In this comparison study, Random Forest achieves a higher prediction accuracy comparing to traditional ARIMA models [22]. It is determined that Random Forest can be used when the dataset has nonlinear relationships or when the model includes additional variables. The findings of the reference paper are applied to the current case study.

In related research by Zhang G.B. from the Neurocomputing Journal, higher prediction accuracy is achieved by using combination of Ensemble of Neural Networks (NN) and traditional ARIMA time series models as outlined in reference [23]. Artificial Neural Networks provide more flexibility than ARIMA models because of nonlinear capability. The hybrid approach of ARIMA with NN allows for the benefits of

nonlinear fitting and also protects against influencing factors such as sample variations and data structure changes. The hybrid ensemble allows for the benefits and strengths of both ARIMA and NN in one model [23]. The findings of the reference paper are considered while building time-series models in current case study.

The study proposed a decomposition method to break down the raw electricity load data into a trend series and a set of fluctuation sub-series data. After the decomposition, researchers are able to apply linear regression model for the trend series data and XGBoost regression model on fluctuation sub-series data. Bayesian optimization algorithm is used to optimized XGBoost hyper-parameters. [24].

Nitesh V. Chawla et al. in reference [25] introduced a new sampling method called Synthetic Minority Over-sampling Technique. The algorithm randomly selects a minority class and find its k nearest neighbors. Then synthetic instance is created by randomly choosing one of the k nearest neighbor. A line segment in the feature space is created by connecting the existing point and newly created instance. This algorithm can be used to create as many synthetic examples for the minority class as necessary. The current study uses XGBoost and it contains weight hyper-parameters which allows us to assign weight coefficient for imbalanced data.

Luo, Zhang, Z. et al. in reference [35] established COVID-19 cases prediction models for the time series data of America by applying XGBoost regression algorithm. Mean absolute error, mean squared error, root mean square error, and mean absolute percentage error are used to evaluate the effect of model performance. By using XGBoost model, a sensitivity analysis was also conducted to determine the feature importance from the model. The current study incorporates the aforementioned approaches used by the authors in dealing with time series data.

The hypothesis here is that the results show the percentage of customers without power can more accurately be predicted using machine learning models and time series methods by leveraging the use of detailed weather data and ERCOT market conditions. Increased prediction accuracy and lower RMSE can be achieved when both feature-sets are included as features in machine learning models or used as exogenous variables in time series analysis.

3. Methods

3.1 Data Source and Database Creation

The data used by the research team was retrieved from multiple sources. The key response variable, outage data, is sourced from 'PowerOutage.us' [26]. 'PowerOutage.us' collects, records, and aggregates live outage data from utilities all over the United States to create the most reliable and complete source of current and historical power outage information [26]. Historical outage data was pulled from August of 2017 through August 2021. This includes the number of electric customers served by county and the respective number of customers experiencing an outage at that time on an hourly basis.

The weather data was retrieved from https://openweathermap.org, a team of research and IT experts that provides historical and real time weather information globally [27]. This includes continuous variables such as temperature, windspeed,

humidity, and precipitation levels and categorical weather descriptions such as cloudy, fog, storm warnings, etc. Current and historical weather information for August 2017 – August 2021 was obtained at an hourly level via the 'OpenWeather' one-call-API.

Historical ERCOT market data was retrieved from the ERCOT online load and pricing archives. Load data by weather zone was available at hourly level. In addition, Real-time and Day ahead pricing at the load zone and the hub were downloaded from the ERCOT archives. The pricing data is at the fifteen-minute interval level and was averaged to show the effective price on an hourly level.

All data was saved and stored as CSV or excel files. Using Python, the 'CSV' files were separately imported into 'Pandas' data frames. Utilizing the 'SQLAlchemy' module, the data frames were then stored into a Microsoft SQL Server Database. The weather data has multiple weather descriptions during the same hour. In such cases, weather descriptions were pivoted to create more columns and limit the data for each hour and county to only one row. The ERCOT data pulled by load zone and weather zone was assigned to the counties that lie within the same geographic region. The data was joined together by county and timestamp, the end result is a data frame containing outage data, weather data, and ERCOT market data.

3.2 Exploratory Data Analysis (EDA)

The entire dataset contains 317,024 rows and 63 columns (See Appendix – Figure B). The first column is the 'datetime' column which records the exact time when the data was pulled from SQL database. The rest of the data features contain time, city, county, weather, electricity prices, temperatures, outage count, customer count and electricity load, etc. By checking the missing values, 'seal_level' and 'grnd_level' attributes don't have any values. Entries 'rain_1h', 'rain_3h', 'snow_1h' and 'snow_3h' have above 86% missing values. Therefore, these columns do not provide any useful information in our future model, which means they are safe to discard.

Before imputing the missing values in the rest of the columns, it is relevant to note that 'ERCOT_WEATHERZONE_LOAD' is not numeric. Therefore, the value type is changed to numeric before imputing the missing values. The missing values are imputed by their own medians—instead of means—to avoid outlier bias. Spaces were replaced with underscore symbol from column names and column values to make them more compatible and usable in data processing and machine learning algorithms. 'RecordDateTime_CST' column was removed from the dataset for the EDA as machine learning algorithms are based on sample independence and date and time records and not included in the feature-set.

In Appendix – Figure C, the temperature plot shows the temperatures from July 2017 to July 2021. Both Houston and Dallas areas have seasonal temperature trends that are consistent with one another. Houston doesn't have extreme low temperatures when compared to Dallas, and both cities had a noticeable temperature drop during February 2021. This was the time when Texas had a winter storm of historic proportion and severity.

In Appendix – Figure D, the outage count plot shows the outage counts from the same period as above temperature plot. The most noticeable is the spike during the Texas winter storm outage, which matches the temperature in Figure C plot. Besides

the winter storm outages spike in both Dallas and Houston, the second highest spike happened around in June 2019 in the Dallas area, and the third highest happened around in September 2017 in the Houston area.

In Appendix – Figure E, the plot shows the ERCOT electricity price of the current load zone. The electricity price had a spike during February 2021. Also, there are spikes in Houston during September 2017, and Dallas during August 2019. Comparing temperature, outage counts, and electricity load price plots, it is not hard to tell that the winter storm in February not only caused electricity prices to rise much higher than any other period in our dataset, but also caused huge electrical outages. However, by taking a closer look at the electricity load prices plot in September 2017 around Houston and August 2019 around Dallas, the severe temperature did not occur during those periods according to temperature plot, but it caused electrical outages in both metroplexes. Looking even closer at the outage plot for Dallas, it appears that the massive outages occurred in June 2019 and electrical load price spike happened in August 2019. The current study aims to identify the important variables and factors that lead to high power outage events.

In Appendix – Figure F, the outage counts corresponding to weather conditions and locations are plotted. Here, snow and freezing rain are selected as extreme weather conditions. During the non-snow days, the average outage counts are ~ 600, but during the snow days, it shows ~ 32,000. Obviously, the winter storm in February 2021 causes the average of power outage to be much higher. In other words, the outages caused by the Texas Winter Storm may be treated as outliers. The data shows that Houston has a high outage percentage during freezing rain conditions, whereas the same weather conditions in Dallas are not correlated with high power outage events. Therefore, it is worth investigating the confounding factors in Houston that led to these higher power outage conditions.

In Appendix – Figure G, the correlation heatmap shows the Megawatt-Hour for ERCOT weather zone is highly correlated with Hub/Load zone energy prices. A couple of temperature features are also highly correlated. Appendix – Figure H shows outage percentages per each major county in Texas. The outliers are easily visualized in boxplot and swam plot. Most of the outliers close to 1.0 are electricity outages happened during Texas Winter Storm in February 2021.

3.3 Data Preparation

To begin to use traditional machine learning algorithms, the dataset is split into modeling group and validation group by county. In each group, the data contains counties in both Dallas-Fort Worth and Houston metropolitan areas. The reason is that by using these modeling and validation set splits, the weather and geographical influences can be minimized. If one group of data only contains counties in a solely metropolitan area, the weather features may have huge differences than the other group. Therefore, 'Harris', 'Tarrant', 'Dallas', 'Montgomery', and 'Brazoria', counties are assigned to modeling group while 'Collin', 'Fort Bend', 'Galveston', and 'Denton', are assigned to validation group.

Since the traditional machine learning models assume sample independence and are not based on serial correlations, all timestamp columns are removed from the dataset. The data is sourced exclusively from Houston and Dallas counties. As such, redundant

geographical columns such as 'State', and 'Metro_Area' are removed from the feature-set. The two Boolean fields that identify records as a 'weekend' and 'weekday' are highly correlated, and only one of them is needed and the other is removed from the dataset. The dataset is strategically split into Modeling and Validation data to be a stratified representation of both Houston and Dallas, as such references to whether the data originated from Dallas or Houston are removed from the feature-set. 'Weekday', 'Month', and 'hour' variables are one hot encoded and changed to categorical variables.

The response variable used in the dataset - 'power_outage_percentage' is created by dividing 'total_outages' by 'customer_count' at each timestamp. With this new response variable in place, 'total_outages' and 'customer_count' are removed from the feature-set. The county column is removed as none of the same counties are present in either validation or modeling set, by leaving this column could create shape errors or training biases in the models. When building classification models, the power outage percentage values are split into four classification buckets. Very low power outages are classified as Class 0 "<1%", small power outages are deemed to be Class 1 "1-3%", medium sized power outages are Class 2 "3-10%" and large power outages are Class 3 "over 10%".

The final step before creating machine learning models is to check whether the values of outage percentage column are all between 0 to 1. If the values are above 1.00, value 1 is re-assigned and if the values are negative, then value 0 is re-assigned. The data was treated with both one hot encoding and normalization to ensure the maximized interpretability of machine learning models. In the regression model section, five-fold cross validation is used with shuffle equals to 'TRUE' and in classification model section, stratified five-fold cross validation is used as it allows each class to maintain the same proportion in all folds.

In this study, two modeling methods are used. The first method uses XGBoost regression results obtained ahead of time to predict the outage percentage and then split predicted outage percentage based on the rule introduced above. The second method splits the outage percentage response variable into four categories before using XGBoost classification algorithm.

For Time series modeling, only continuous data features can be used. As such, the feature-set was stripped down to only include weather features: temperature, windspeed, humidity and rainfall. In addition, ERCOT features: load, real time load zone and hub prices, and day ahead load zone and hub prices. Unlike traditional machine learning models, time series depends on serial correlation to previous known response variable values. As such, hourly outage percentage was included in the feature-set as well. The full dataset currently has time series data on nine different counties. For purposes of model building only two counties were selected. Dallas county is used as the training set and Tarrant County as the test/validation set. The realizations were trimmed to only use time series data from Sept 1 2019 – August 31 of 2021 (two years). To accommodate for previously known values, additional columns for each feature were created based on their lagged time series value. The full time series feature-set has the current and lagged hourly values (up to last 15 hours) for each of the variables in the feature-set.

3.4 Linear and non-linear Regression Models

This section of machine learning modeling uses the following linear regression models: lasso (L1), Ridge (L2) and Huber regression along with non-linear regression models including random forest, gradient boost, decision tree, support vector machine, and multiple layer perceptron regression.

Lasso and Ridge both solve the linear least square loss function with regularization applied to penalize their regression coefficients to avoid overfitting the problem. Lasso uses the absolute value of the coefficients and Ridge uses the square value of the coefficients. Lasso introduces sparsity, which can reduce the feature coefficients to zero for feature selections, while Ridge regression does not. Ridge regression still allows features to contribute even if this feature does not contribute much to the model. The Huber regressor is robust to outliers because its loss function is a balanced comprised between squared loss which is centered around the mean and absolute value loss which is centered around the median [36]. In this paper, strength of the penalty 'λ' is tuned to minimize the RMSE of the three linear regression models. When $\lambda = .0001$, lasso regression yields the lowest RMSE out of fold loss and when $\lambda = .001$, Ridge and Huber regression yield the lowest RMSE out of the fold loss.

Random forest and gradient boost both belong to ensemble algorithms. They gain advantages by combing several basic estimators to build a given learning algorithm so as to improve the accuracy from a single estimator. The different ensemble methods used between random forest and gradient boost is that random forest builds several estimators independently to reduce their variance. Gradient boost builds base estimator sequentially to reduce the bias of the combined estimator [37].

Random forest and gradient boost also belong to decision tree family. The algorithms predict response variables by learning decision rules inferred from the data features. Random forest fits a couple of classifying decision trees on various sub-samples, then averages the prediction accuracy and controls the over-fitting. Gradient boost is built in a stage-wise fashion, like other boosting methods, but it allows to optimize arbitrary differentiable loss functions [38][39].

Support vector machine (SVM) uses the kernel trick and then maps the outputs from kernel functions into high dimension feature spaces. Different kernel functions can be specified for the decision function in a single prediction problem. Therefore, SVM is also useful on unlabeled data to find the natural clustering of the data and map the data to hyper-plane to categorize the data labels [40].

Multi-layer perceptron (MLP) is one of the neutral network algorithms. The first layer comes from all the features in the data, and they are represented by a set of neurons. Starting from the second layer, all the layers are called hidden layers. Each neuron in the hidden layers transforms the values from the previous layers with a weighted linear summation followed by a non-linear activation function [41].

3.5 XGBoost

The XGBoost algorithm is a convenient algorithm which can be used either in classification or regression. Researchers can tune its hyper parameters to deal with missing values, down-sample the data size, random select features in the data to avoid overfitting, and provide the weight of the data to handle unbalanced dataset. It is

important to mention that XGBoost belongs to one of the tree algorithms. Thus, researchers don't have to normalize the data before using this algorithm. Multi-collinearity is also non-existent among data features by using tree-based model. Therefore, this paper heavily relies on XGBoost algorithm to achieve the best regression and classification results. The basic functionality behind XGBoost algorithm is shown in Fig. 2. XGBoost also provides various objective solvers for different types of problems. In this paper, 'reg:squarederror' is used to solve regression problem and 'multi:softprob' is used to solve multi-class classification problem. Another useful feature in XGBoost is 'scale_pos_weight' hyper-parameter, which helps to assign the weight coefficient to one of the classes in binary classification problem when imbalanced data is introduced. But 'scale_pos_weight' only works well in binary classification and XGBoost doesn't provide other scale weight attribute to tune for multi-classification problem. Therefore, in this paper, it is a necessary and significant step to calculate the penalization coefficients for each class and balance the sample weight before fitting the model. In conclusion, predicting multi-classification is to predict the probability of each class.

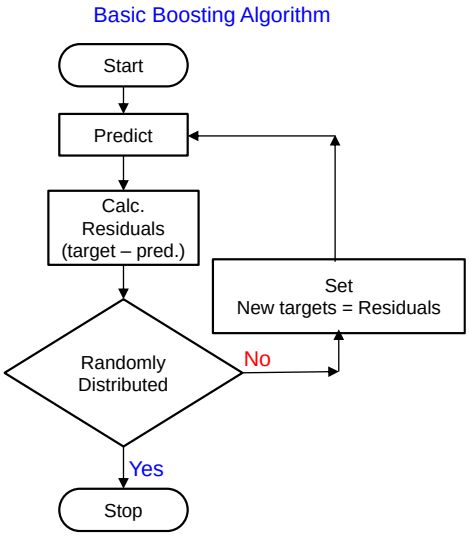


Fig. 2. XGBoost logic flow [24].

3.6 Time Series Models

In addition to traditional machine learning models that assumed independence of observations, the data was fitted with various time series models that assume serial correlation. Serial correlation (or autocorrelation) refers to the similarity between observations as a function of the time lag between them [29]. The following models are based on this assumed relationship.

To determine the appropriate number of lagged observations to include in the time series models, the AIC (Akaike Information Criterion) and BIC (Bayesian Information

Criterion) were taken for each autoregressive model of 1 through 15 lags on the training data. The number of lags that report the lowest AIC/BIC are used for model building.

Given the large dataset and number of realizations, only a subset of the full data is used for time series model building. For this time series analysis, only two years of hourly data is used (Sept 2019 – August 2021) with Dallas County as the training set and Tarrant County as the test/validation set. With the sparsity of the data, outage percentages are broken out into four classes. The majority of outages percentages fall under 1% (99%). Special consideration is given to how model performance in other classes, especially when outage % are high ("3-10%" and "10% +").

Table 1. Outage percentage category by county.

					Outage % Bucket			
Date Range	# of Hourly Records	County			< 1%	1-3%	3-10%	10% +
9/1/2019 - 8/31/2021	17,520	Training Set	Dallas	Count:	17,308	102	53	57
				% of Total	99%	0.58%	0.30%	0.33%
		Test Set	Tarrant	Count:	17,368	60	20	72
				% of Total	99%	0.34%	0.11%	0.41%

To assess the accuracy of time series models, the RMSE and the Balanced class Recall are measured in a 24-hour rolling window. In the rolling window forecast, the first hour is predicted using all actual lagged features. As the window moves along, most recent lagged features are replaced with the of the prior forecasted outage value.

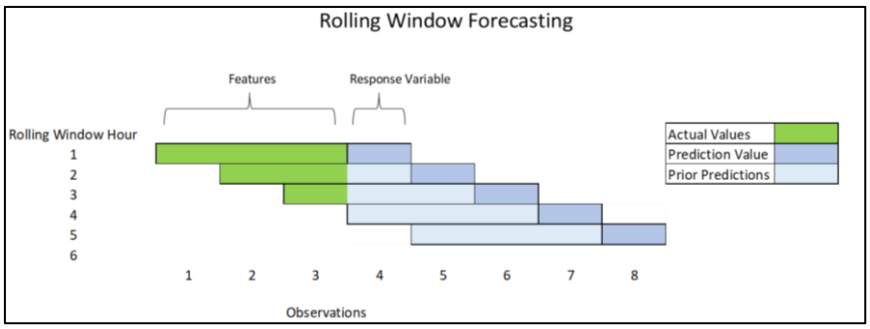


Fig. 3. Time Series Rolling Window Forecasting.

As the window moves along, predictions are based solely off the value of prior predictions. When the window is complete, the next rolling window is built starting from the next observation in the test dataset. The rolling window method shows how well each model is in predicting the power outage percentage over the next 24 hours. Given the two years of hourly data in question (17,520 observations) and the rolling window size of 24 hours allows for 17,496 rolling windows. The rolling windows are tested, by taking 24 hourly predictions at each of the 17,496 rolling window positions, and a total of 419,904 predictions are obtained. The RMSE and balanced recall results for each hour are shown in Fig. 6. through Fig. 12. to determine the best fit of each model.

In addition to time series models, several control groups were also tested to ensure the accuracy of fitted time series models. The control groups are the average outage percentage found in the Dallas County dataset (approx. 0.12%) and what the last known outage % was at the beginning of each rolling window. These are held constant in each rolling window vs. the results of the fitted model. This check ensures that models have lower RMSE and higher Recall than simple control methods involving no time series methodologies. The Fig. 4 below outlines how the control groups are utilized in the rolling window vs fitted models.

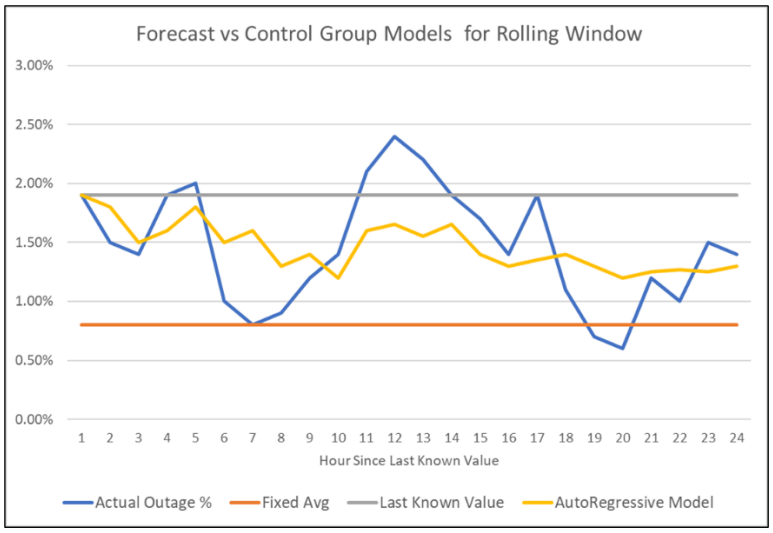


Fig. 4. Forecast vs. Control Group models for rolling window

Initial models that are fit were univariate models including linear based autoregression, random forest, and neural networks. The autoregression models are trained on Dallas data using OLS (Ordinary least squares) to find the best fit. The random forest and neural network models were tuned using five-fold class stratified cross validation to find the highest performing model.

Multivariate time series models were built using linearly based Vector Autoregression (VAR). All combinations of feature-sets were tested: i) outage and weather features only, ii) outage and ERCOT features only and the combination of all three feature-sets iii) outage, weather and ERCOT features. VAR was also tested using different combinations of data feature quality. Models were initially tested using the actual values of data features, the last known feature values at the start of a rolling window, and different prediction methods for features. Variables were treated as either endogenous (where they are explained by other variables in the model) or exogenous (variables not explained by other variables in the model). Endogenous forecast was done using VAR for each feature where all other variables are included in the prediction of each other within the rolling window. By comparison, the exogenous VAR forecast was done by building separate univariate linear auto regressive and univariate neural networks to predict feature values within the rolling window.

Table 2 below demonstrates the use and explanation of actuals, last known values and forecasted features with exogenous & endogenous forecasting methods. As the model traverses the rolling window, Table 2. establishes how the predicted y values for each model method are plugged back into the feature set for each subsequent rolling hour.

Table 2. Actual, Last Known value and Forecast Values

		Models with Actuals			
		Prediction Observation			
		6	7	8	9
X -Variable					
Lag 1	→	5	6	7	8
Lag 2	→	4	5	6	7
Lag 3	→	3	4	5	6
Lag 4	→	2	3	4	5
Lag 5	→	1	2	3	4

		Models with Last Known Value			
		Prediction Observation			
		6	7	8	9
X -Variable					
Lag 1	→	5	5	5	5
Lag 2	→	4	5	5	5
Lag 3	→	3	4	5	5
Lag 4	→	2	3	4	5
Lag 5	→	1	2	3	4

		Models with Forecast Value			
		Prediction Observation			
		6	7	8	9
X -Variable					
Lag 1	→	5	Forecast	Forecast	Forecast
Lag 2	→	4	5	Forecast	Forecast
Lag 3	→	3	4	5	Forecast
Lag 4	→	2	3	4	5
Lag 5	→	1	2	3	4

The model with top performing RMSE and balanced recall at the later lags of the 24-hour rolling window are selected for further analysis. The class level precision scores and recall scores are examined at the hourly level.

3.7 Web Application

The final output of the study is an interactive online web application. The best fit machine learning XGBoost model is exported from Python using the 'Pickle' and 'Joblib' functions. The model and Python functions are embedded in a 'pywebio' application [42] that allows for web hosting and public access. The user interface in the

pywebio application has a drop-down menu to select Texas counties. Once a county is selected, the application pulls necessary weather features from the 'OpenWeather' API and the 'current price', 'day ahead price', and load data from ERCOT are screen scraped from the current ERCOT market conditions website [29]. These datapoints are converted into a Pandas dataframe and fed into the imported machine learning model. The predicted 'power outage percentage' based on real time actual feature inputs is displayed on the web application. The results are displayed against current power outage statistics pulled via 'PowerOutage.us' [26]. The schematic for the entire web application process is shown in Fig 5 below.

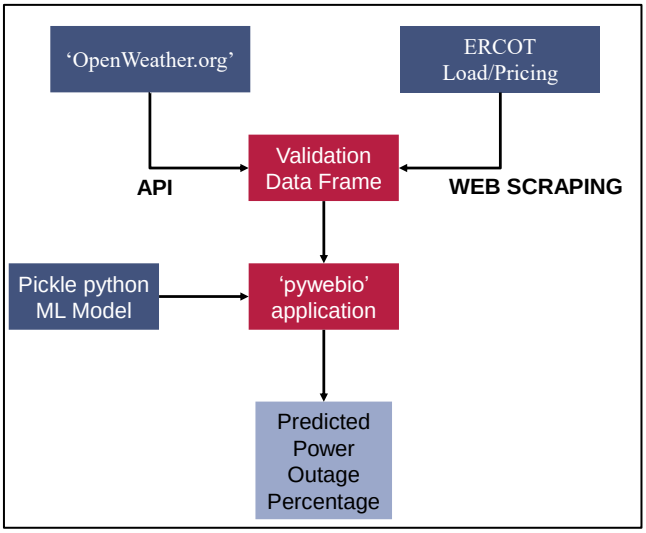


Fig. 5. 'PYWEBIO' web application process

4 Results

The results presented in the below sections can be considered as extension of the historical data analysis, deterministic simulation, and high-level statistical models discussed Andersson G et al. in reference [1]. The current ERCOT analysis results can be addition to various studies by Carreras B. A. et al. in reference [2] for the Eastern and Western interconnections of the North American (NERC) grid. The authors in reference [2] present the factors that influence the power outages, statistical quantification and time correlations; the time series studies in current study can supplement these former studies. Ji, C. et al. in reference [6] determines that extreme weather is not the only cause for power outages, but rather exacerbates, existing vulnerabilities that are obscured in daily operations. The current study results also boost the reference [6] results that extreme weather is not the only cause of power supply issues.

4.1 Machine Learning Models

As described in the previous sections the entire dataset is split into modeling and validation groups by county. Modeling group includes Tarrant (DFW), Dallas (DFW), Montgomery (Hou), Brazoria (Hou), Harris (Hou). Validation group includes Collin (DFW), Fort Bend (DFW), Galveston (Hou), Denton (DFW).

When building classification models, the power outage percentage values are split into four classification buckets. Very low power outages are classified as Class 0 "<1%", small power outages are deemed to be Class 1 "1-3%", medium sized power outages are Class 2 "3-10%" and large power outages are Class 3 "over 10%".

Table 3. shown below lists regression RMSE from traditional machine learning models, RMSE is the average of five-fold cross validation values yielded by the designated regression algorithms applied to modeling data group.

Table 3. Regression RMSE for Traditional ML models
(including both ERCOT and Weather features)

Type	Algorithm	RMSE
Regression	Lasso	1.46%
Regression	Ridge	1.46%
Regression	Huber	1.74%
Regression	Random Forest	1.20%
Regression	Gradient Boost	1.17%
Regression	Decision Tree	1.09%
Regression	SVM	1.60%
Regression	MLP	1.19%
Regression	XGBoost	.78%

XGBoost yielded the best RMSE without any doubt. Therefore, this paper uses XGBoost regression to plot the feature importance. (See Appendix - Fig I. 'Weekday_3', ERCOT load zone price and 'Clear_sky' are top three features obtained by regression model.)

Table 4. shown below lists the regression results by XGBoost with weather features removed or ERCOT load prices removed.

Table 4. XGBoost with ERCOT prices and Weather only models

Type	Algorithm	Features	RMSE
Regression	XGBoost	ERCOT load prices only	1.45%
Regression	XGBoost	Weather only	.93%

The results show that machine learning models that utilize both weather features and ERCOT market data yield a lower RMSE than using one feature-set exclusively.

Table 5. below shows the RMSE of the XGBoost model using weather and ERCOT features on the validation set. The model was trained on the modeling data and independently tested on the validation data.

Table 5. XGBoost Validation model results

Type	Algorithm	Data group	RMSE
Regression	XGBoost	Validation	1.04%

Table 6. below shows the classification performance for Class 3 ("over 10%") by using regression XGBoost model prediction results. Class 3 is the group with power outage percentages larger than 10%.

Table 6. XGBoost Confusion Matrix – 100% Validation group data

Algorithm	Precision Class3	Recall Class3	F1 Score Class3
XGBoost	77%	94%	85%

Confusion matrix

Actual / Predict	0 (<1%)	1 (1% to 3%)	2 (3% to 10%)	3 (>10%)
0(<1%)	137464	705	174	27
1(1% to 3%)	761	75	39	16
2 (3% to 10%)	324	84	77	28
3 (>10%)	8	4	4	241

Table 7. below shows the classification performance for Class 3 by using classification XGBoost model. Class 3 shows the group power outage percentages larger than 10%.

Table 7. XGBoost Confusion Matrix – Modeling group data (20% test set)

Algorithm	Precision Class3	Recall Class3	F1 Score Class3
XGBoost	70%	89%	78%

Confusion matrix

Actual / Predict	0 (<1%)	1 (1% to 3%)	2 (3% to 10%)	3 (>10%)
0(<1%)	31291	2984	461	37
1(1% to 3%)	75	159	25	0
2 (3% to 10%)	8	28	58	1
3 (>10%)	1	7	3	89

Please note that the total counts of Table 6 and Table 7 confusion matrices are different because the first classification in Table 6 used entire validation group data and the second classification in Table 7 only used test dataset which is 20% of modeling group data.

Appendix – Figure J shows train and test Receiver Operating Characteristic plot of XGBoost Classification. The training ROC is above test ROC all times which means the model didn't overfit the data.

Appendix – Figure K shows Receiver Operating Characteristic plot for all classes. Outage percentage larger than 10% group (Class 3) has the highest AUC.

Appendix – Figure L shows feature importance from XGBoost classification. The top three features are ERCOT load zone prices, 'month_8' and 'pressure' which ERCOT load zone prices are still one of the top three important features.

4.2 Time Series Models

After testing all feature-sets on all combination of lags (1 through 15), lag 15 produces the lowest AIC and BIC, as such all-time series models are built as a function of the prior 15 hourly observations. The results for AIC and BIC are shown in Table 8. Below.

Table 8. AIC and BIC for Time Series models

feature_set	lags	aic	bic	AIC Rank	BIC Rank
outage_only	15	-11364	-11248	1	1
outage_weather	15	-11513	-10961	1	1
outage_ercot	15	-12841	-12398	1	1
outage_weather_ercot	15	-13010	-12132	1	1

Fig. 6. and Fig. 7 below show the result of the Univariate vs. Control Groups time series for RMSE and Balanced class Recall.

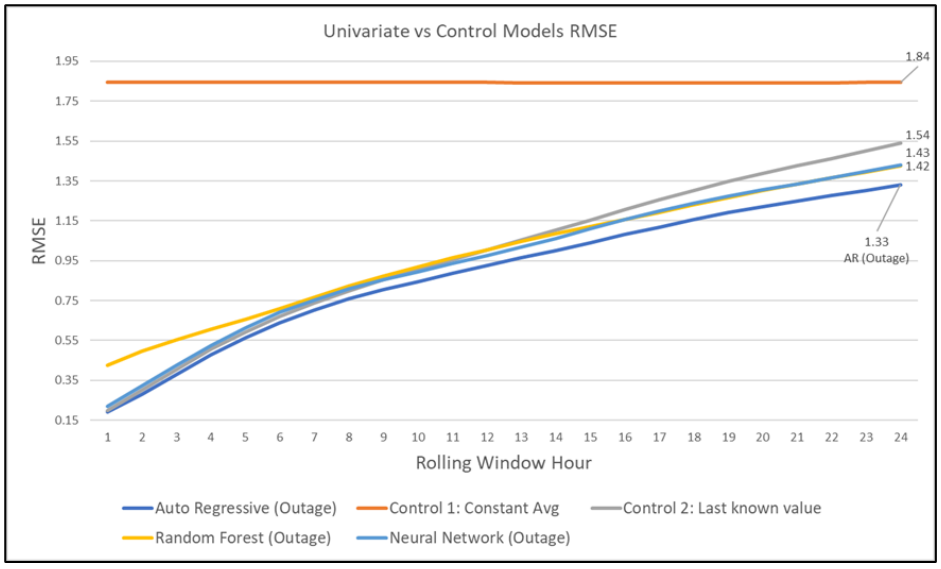


Fig. 6. Univariate vs. Control Group RMSE

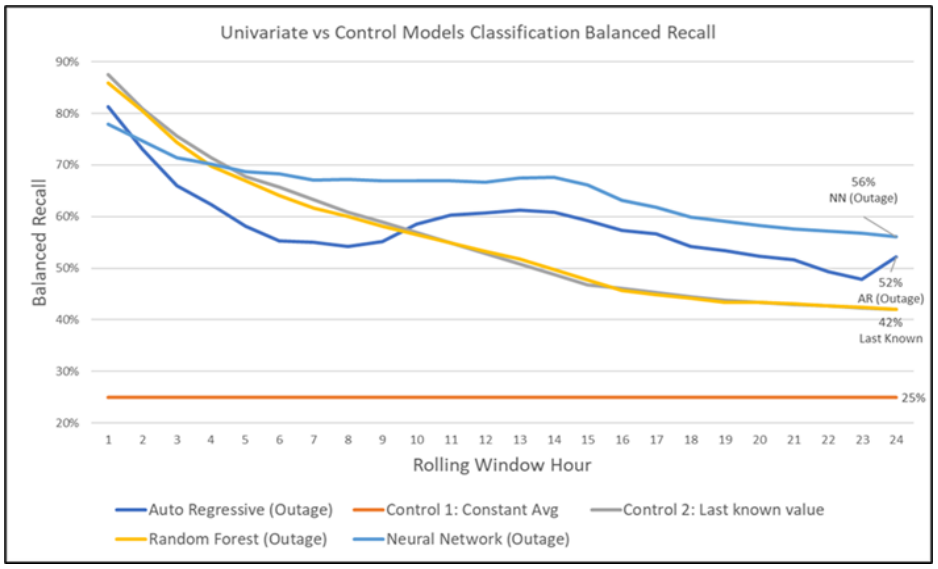


Fig. 7. Univariate vs. Control Group Balanced Recall

All univariate models outperformed the control groups in both RMSE and Recall. The Autoregressive Linear model maintained the lowest RMSE (1.33) at 24 hours vs all other models whereas the univariate neural network has the highest balanced class accuracy at 24 hours (56%).

The following (Fig. 8. and Fig. 9.) show how VAR perform using different combination of feature-sets: outage & weather, outage & ERCOT, outage with weather & ERCOT using the actual values of the non-outage variables.

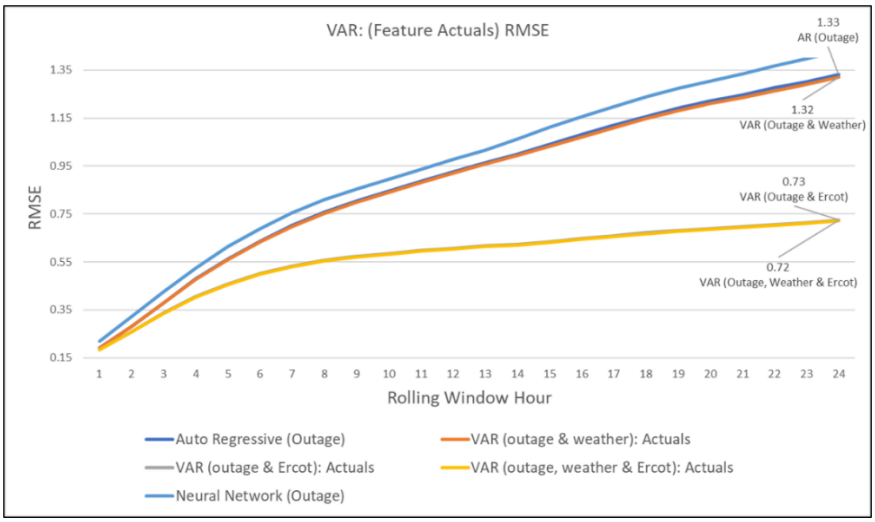


Fig. 8. VAR RMSE with different combination of features.

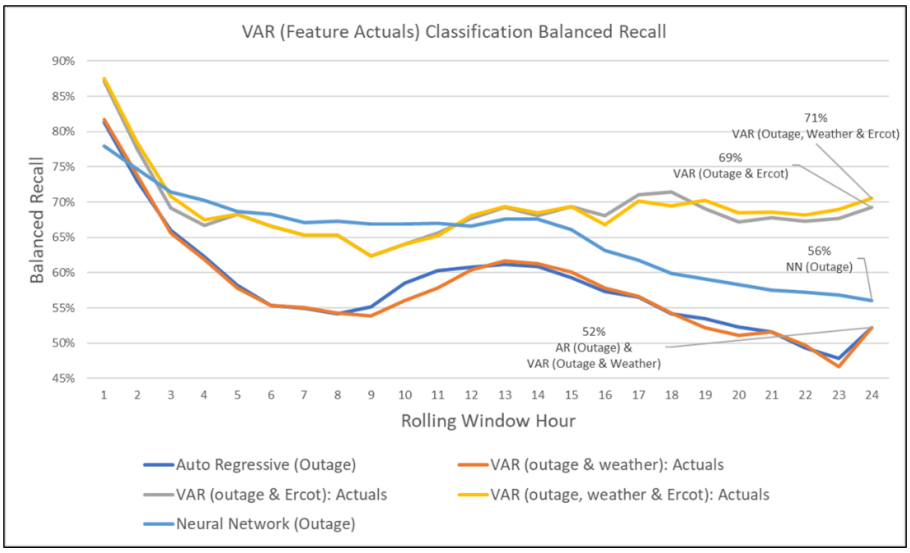


Fig. 9. VAR Balanced Recall with different combination of features.

VAR models greatly outperformed both AR & NN models in both RMSE and Recall. The VAR using actual features for outage, weather, & ERCOT has the lowest RMSE (.72 at 24 hours) and highest Recall Score (71% at 24 hours).

Fig. 10. and Fig. 11. show how VAR performs using different combinations of features. Models were built using actual known features, last known features prior to the 24-hour window, and forecasted features by either univariate autoregression (exogenous) or multivariate vector autoregression (endogenous).

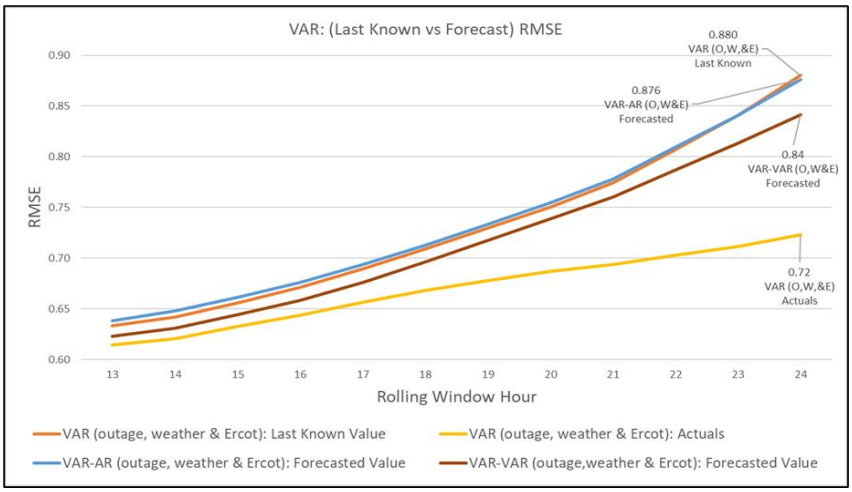


Fig. 10. VAR RMSE (Actual vs. Last Known) with different combination of features.

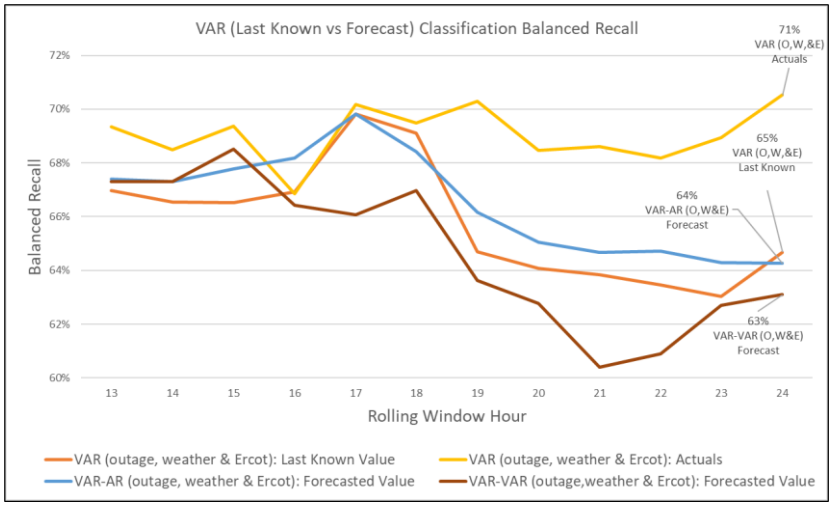


Fig. 11. VAR Balanced Recall (Actual vs. Last Known) with different combination of features.

The VAR model using the full feature of outage, weather and ERCOT and exogenously forecasting weather and ERCOT features by univariate autoregressive models has the lowest RMSE (.72 at 24 hours). In addition, it maintained the highest consistent recall score for rolling window hours 19 through 24 (64% at 24 hours).

Fig. 12. and Fig. 13. show the class level precision and recall scores for the VAR model using the full feature of outage, weather and ERCOT and exogenously forecasting weather and ERCOT features by univariate autoregressive models.

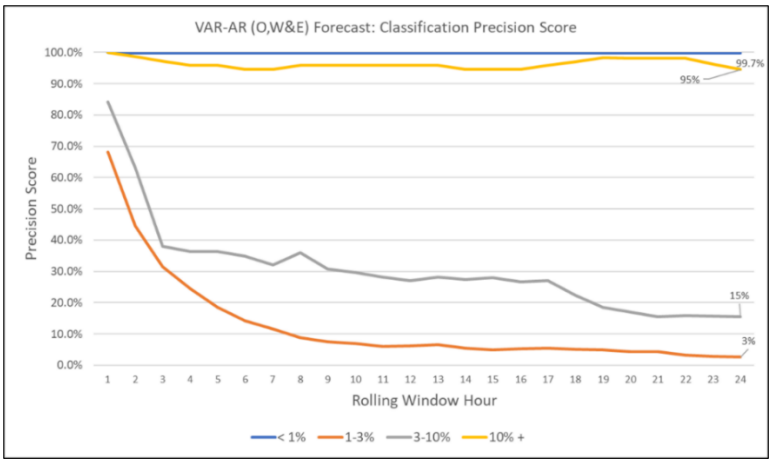


Fig. 12. VAR - AR Precision Score with classification using all features.

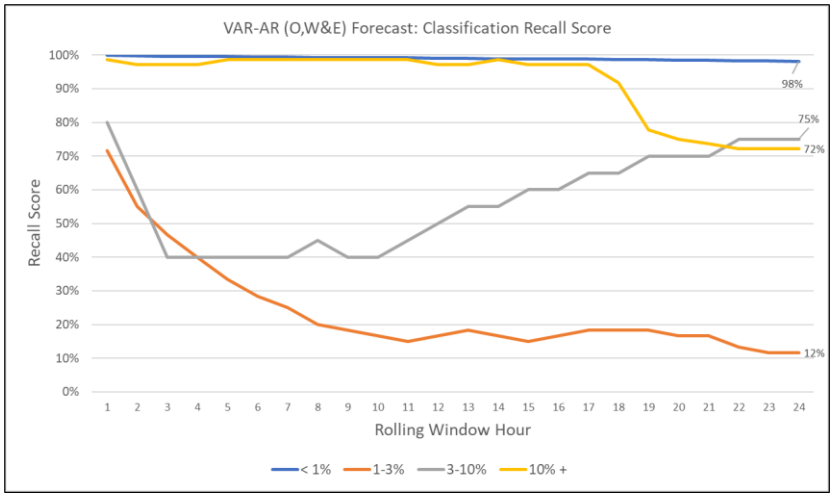


Fig. 13. VAR - AR Recall Score with classification using all features

The model maintains 99% precision and recall for the most common class (<1%). Moreover, high precision and recall scores for high-risk outage buckets (95% at hour 24 precision for 10%+ bucket and 75% / 72% at hour 24 Recall for 3-10% and 10%+ buckets respectively. Admittedly, the model has low precision for 1-3% and 3-10% buckets, 3% and 15% respectively.

Table 9. summarizes the overall results of time series models across every hour and every rolling window. The RMSE, R squared value and balanced recall for each model are shown.

Table 9. RMSE, R^2 and Balanced Class Recall for Time Series Models

Model No.	Model Fitted	RMSE	R2	Outage Bucket: Balanced Class Recall
1	Control 1 -fixed value	1.84	0.000	25.0%
2	Control 2- Last known Outage %	1.06	0.671	55.7%
3	Random Forest (Outage)	1.03	0.687	55.3%
4	Neural Network (Outage)	1.01	0.697	65.2%
5	Auto Regressive (Outage)	0.95	0.735	58.2%
6	VAR (Outage & Weather) Actuals	0.94	0.739	57.9%
7	VAR (Outage & Ercot) Actuals	0.59	0.899	68.8%
8	VAR (Outage, Weather & Ercot) Actuals	0.58	0.900	69.1%
9	VAR (Outage & Weather) Last Known	0.95	0.737	58.0%
10	VAR (Outage & Ercot) Last Known	0.62	0.885	67.2%
11	VAR (Outage, Weather & Ercot) Last Known	0.63	0.885	67.4%
12	VAR (Outage & Ercot) AR Forecasted	0.63	0.884	67.7%
13	VAR (Outage, Weather & Ercot) AR Forecasted	0.63	0.884	67.8%
14	VAR (Outage & Ercot) NN Forecasted	1.49	0.35	58.4%
15	VAR (Outage & Ercot) VAR Forecasted	0.61	0.889	66.9%
16	VAR (Outage, Weather & Ercot) VAR Forecasted	0.61	0.889	67.0%

4.3 Web Application

The deployed web application is shown in Fig. 14. The Fig. 14. displays highlighted boxes which are described as follows: Highlighted box ① indicates the header for the app, box ② shows the selection drop down list for the 9 counties in Dallas – Fort Worth and Houston areas. Once a selection is done, the current weather conditions for the selected county is shown in box ③ and the county map highlighted on Texas state map in box ④. The outage conditions for selected county are displayed as shown box ⑤ and the ERCOT market conditions which are used in prediction are shown in box ⑥. The final result – predicted outage percentage is shown box ⑦.

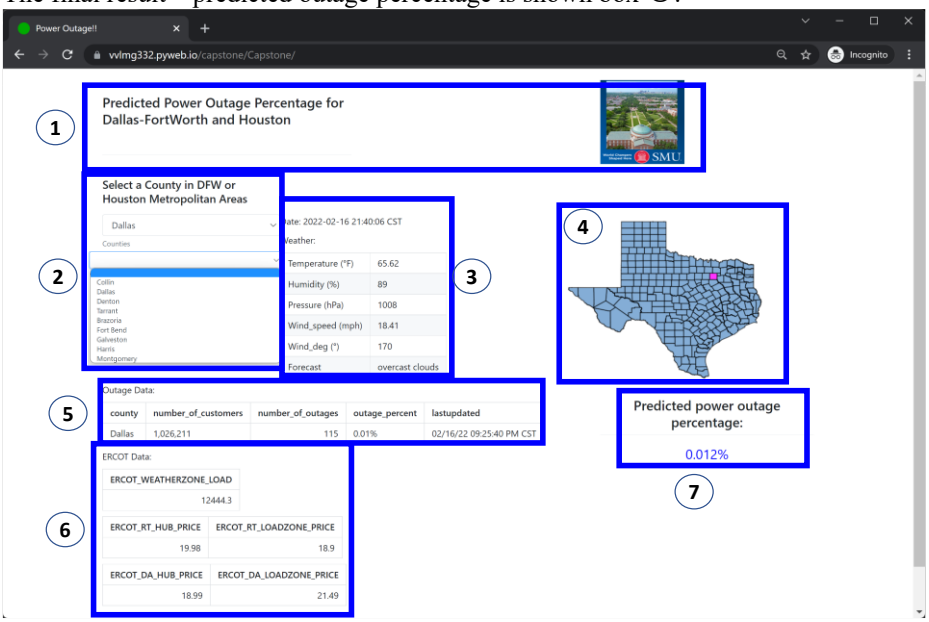


Fig. 14. Pywebio Deployed Output

5 Discussion

5.1 Results Discussion

Linear and non-linear regression models including XGBoost, Lasso, Ridge, Random Forests, Gradient Boosting, Neural Networks, SVM and Decision Trees are trained and tested to predict the power outage percentage for each of the selected Texas counties using weather data and ERCOT market conditions. XGBoost outperformed the other algorithms tested. When assessing the individual fold level results from five-fold cross validation in all models tested, there is always one-fold RMSE value that deviates away from the RMSE of the other four folds. After plotting the response variable outage percentage distribution and zooming it in, it is clear that the outage percentage distribution is heavily skewed right (Appendix – Fig. M.). Carefully investigating the

data, the highest power outage percentages happened during the February 2021 Texas Winter Storm and the second highest power outage percentages happened during the Hurricane Harvey from August 2017 to September 2017.

Both XGBoost regression and XGboost classification models indicate ERCOT load zone prices contribute the most important features rather than severe weather. It is observed that bad weather conditions do rank high in feature importance plots, but not as high as ERCOT load zone prices in either model. Therefore, ERCOT load prices play the key feature in the power outage prediction study. A further study could be conducted from the regions of Texas outside of the Houston and Dallas counties, as it is questionable that load zone prices in these less densely populated regions still contribute to power outages in the same way.

Back to the models themselves, what is interesting is that the XGBoost regression performs better than XGBoost classification. One main reason is the data is heavily skewed. Though equal sample weights have been assigned to each class during the model training process; evaluation AUC score was only evaluated by probability of one class (Outage percentage larger than 10%), not across all the classes. During the regression, the evaluation metric RMSE was optimized across all the response variables. In the power outage percentage classification model, the algorithm predicts each class with the highest prediction score. But under the hood, the metric calculates each class prediction probability by grouping other classes into a second class. Then this 'binary' metric is averaged over all classes to get either a weighted average or macro average scores. Therefore, in the future study, a threshold for a typical class can be tweaked based on what predicted answers researchers can accept. To make it easier in the future, binary classification also could be conducted directly, and its performance could be better than regression model results – but it is highly dependent on how researchers split their classification categories.

Time series modeling showed how even univariate time series models such as linear autoregressive, random forest, and neural networks can drastically help improve power outage percentage prediction metrics versus the control groups that assume the flat average or the last known power outage percentage for the next 24 hours. It was found that inclusion of continuous variables from weather data and ERCOT help further drive lower RMSE and Recall Score. Similar to machine learning models, inclusion of the ERCOT data greatly increased prediction accuracy versus that of univariate models. Inclusion of all three (weather, ERCOT & outage) provides a more modest increase than just outage data and ERCOT alone. The model with the highest accuracy used actual data points for weather and ERCOT data, although in a future forecast setting this is not possible as these values will be unknown. By assuming the weather and ERCOT features are exogenous and using individual linear autoregressive models to predict each feature as it traverses the rolling window, the research team was able to achieve lower RMSE and higher recall score than by just assuming the last known values at the window start. Time series modeling showed high recall scores for high-risk outages classes “3-10%” and “>10%”. On the other hand, it showed very low recall scores for the “1-3%” class. This could imply that the model can only predict systematic outages and not those on a localized or neighborhood level. Smaller specific neighborhood issues such as a transmission line break are not picked up by the model or the feature-set. The high recall scores for high-risk outages classes make this model reliable. And because false positives for high outage percentage do not carry any

inherent risks or consequences, it is much safer for consumers if the model provides several false positives on when outage risks are high (low precision) as long as it captures when the majority of high-risk outage events actually occur (high recall).

With validation of research results, models can be extended to rest of Texas ERCOT regions to make power outage predictions. The accuracy of predicting the power outages helps the consumers be aware of the risk and be prepared to handle the power outage situations. The results can only be inferred inside the ERCOT electrical grid system. It cannot draw statistical inference to the remaining regions of Texas which are not under ERCOT, or other states in the U.S. where there is not a comparable market structure, historical data, or weather patterns.

5.2 Ethics Discussion

Considering the ethical attributes, while the prediction results for high-risk outage classes were dependable, the model should not be used by consumers to make decisions that could affect their health or have life or death consequences. Causal inferences between ERCOT loading prices and power outages cannot be drawn. The research team advises end users to take all necessary and reasonable precautions in case of power outage, regardless of the prediction models.

In addition, the results of this model could be used in bad faith. If the model predicts possible power outage risks, this could lead to product hoarding or price gouging of supplies that could put more consumers at risk. This could also extend to energy traders, power generators, or any other market stakeholders in ERCOT that could use these results for monetary gain instead of spreading safety awareness. Some stakeholders may take preventive action, diminishing the usefulness of the model. Larger commercial and industrial customers or cooperative/municipalities can have backup generators, batteries for energy storage, or solar panels at their disposal where the need for knowing power outage risk becomes lower as business operations are uninterrupted by grid failure.

The model predicts outages based on the demand of the current population in Houston, Dallas and the rest of Texas, so any change in growth over time impacts the ability to predict. According to the US Census Bureau, Texas has had the nation's largest annual population growth every year from 2010 through 2016 [44]. Furthermore, Dallas, Houston, Austin and San Antonio lead the population growth in the state [44]. While recent history is concerning, the population gains are expected to grow. An estimate published by the Texas Water Development Board expects population increases in Dallas County by 10% in 2030 and 22% 2040 [45]. There are similar projections for Houston showing population percentage gains of 7% and 14% in 2030 and 2040 [45]. Without additional interconnects to other NERC regions, the Texas power grid will be under even more strain in the years to come. The model should be retuned and retrained as the market and regulations change and evolve. Changes in infrastructure relating to additional transmission lines, cold weather protection investment, gas pipelines and renewable energy will have undetermined effects on the future reliability of ERCOT. There could also be changes in regulation or policy that change the meaning, significance and value distributions of variables utilized by the model. New policies such as price caps or capacity payments would need be recorded and evaluated into model performance to maintain accurate results.

The way that ERCOT variables are used in machine learning and time series models may also be paradoxical. While high load and high prices were identified as influential variables, these can be deceptive because of a feedback loop for how the ERCOT market operates. As the prices go up, generator resources are more likely to come online because of economic incentive, the prices would quickly return to nominal levels. From these variables alone, it is unclear how long the market can operate before load shedding actions must be taken and mandatory blackouts be implemented. The model may show a high risk of power outages for a specific instantaneous moment that quickly goes away because generation resources were promptly dispatched. While the risk for power outages at that time was valid, end users may unnecessarily be alarmed or get prepared for an issue that gets quickly and naturally resolved by the ERCOT market.

The interpretation of weather patterns and its effect on the grid outages in Texas need to be examined as well. Events such as freezing can cause outage from higher demand (perhaps from electric heaters or other appliances) as well as down power lines from frozen rain or snow. Excessive heat can lead to higher demand (from air conditioning and other appliances) but also overload specific power lines. Extreme weather conditions can cause either systematic power supply issues or outages felt at a more localized and neighborhood level. Extreme weather will not always lead to higher demand or decreased generation supply, as such these conditions will not always imply a high risk of power outages per the model. Events felt on local level caused by storms will not affect all residents of large metro areas like Dallas and Houston. Weather only appears for major outages when it correlates and corresponds to high-risk activity from ERCOT load and pricing.

An alternative outlook of power outage causes has been identified and investigated. Malfunction of the power generation equipment or severe weather may never be solely responsible for widely spread power outages. Momentary drastic increases in pricing or load can get quickly addressed and rectified without consumers ever knowing. The ERCOT power grid is complex with many stakeholders, variables and moving pieces. The predictions and inferences made from this model should be taken with caution to ensure the preparedness and safety of its end users.

6 Conclusion

The goal of this study is to build machine learning and time series models that can accurately predict the current risk of power outages as well as the predicted hourly risk over the next 24 hours. While not outwardly apparent, ERCOT features such as Load, Real Time, and Day ahead prices are most influential in determining power outage risk, and accuracy is maximized when done in consideration with weather data. When assuming sample independence and non-serial correlation between records, XGBoost produces the most accurate results. In comparison, in time series modeling, using Vector autoregression with the full feature-sets of outages, weather and ERCOT and exogenously forecasting weather and ERCOT features by univariate autoregressive models produces the most accurate model for forecasting the next 24 hours. Both XGBoost and VAR utilizing the full feature-set yielded a lower Root Mean Squared Error and higher Recall score than using either feature-set exclusively.

Using these feature-sets allows for immediate prediction and forecasting results as current values are freely available by either screen scraping from ERCOT or API for

'Openweather'. The top performing machine learning model is packaged into an external facing web application for public use in determining current power outage risk. This research has provided machine learning and time series models that can be further extended to entire Texas ERCOT regions based on the availability of data.

Acknowledgments. Jake Drew, PhD. – Capstone Professor

References

1. Andersson G., Donalck P., Farmer R., Hatziargyriou N., Kamwa I., Kundur P., Martins N., Paserba J., Pourbeik P., Sanchez-Gasca J., Schulz R., Stankovic A., Taylor C. and Vittal V., Nov. 2005. Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance, in IEEE Transactions on Power Systems, vol. 20, no. 4, pp. 1922-1928, doi: 10.1109/TPWRS.2005.857942.
2. Carreras B. A., Newman D. E., and Dobson I., Nov. 2016. North American Blackout Time Series Statistics and Implications for Blackout Risk, in IEEE Transactions on Power Systems, vol. 31, no. 6, pp. 4406-4414, doi:10.1109/TPWRS.2015.2510627.
3. Witthaut D., and Timme M., (2012). Braess's paradox in oscillator networks, desynchronization and power outage. New Journal of Physics. 14. 083036. 10.1088/1367-2630/14/8/083036.
4. Pahwa S., Scoglio C.M., & Scala A., (2014). Abruptness of Cascade Failures in Power Grids. Scientific Reports, 4.
5. Vaiman M., Bell K., Chen Y., Chowdhury B., Dobson I., Hines P., Papic M., Miller S., and Zhang P., (2012). Risk Assessment of Cascading Outages: Methodologies and Challenges, in IEEE Transactions on Power Systems, vol. 27, no. 2, pp. 631-641, May, doi: 10.1109/TPWRS.2011.2177868.
6. Ji, C., Wei, Y., Mei, H. et al., (2016). Large-scale data analysis of power grid resilience across multiple US service regions. Nat Energy 1, 16052.
7. NERC. (n.d). North American Electric Reliability Corporation https://www.nerc.com/news/Pages/LearnAboutNERC.aspx.
8. Ercot. (n.d.). Electric Reliability Council of Texas. http://www.ercot.com.
9. ERCOT's market structure and oversight. http://www.ercot.com/content/wcm/lists/190192/Market_Structure_OnePager_FINAL_Revised.pdf.
10. Bureau, US Census. "Metropolitan and Micropolitan Statistical Areas Totals: 2010-2019". The United States Census Bureau.
11. "Census finds 4 million new Texans, enough for 2 extra US House seats, though we expected more". Dallas News. 2021-04-26.
12. Ura, Alexa (2021-04-26). "Texas will gain two seats in Congress as residents of color drive population gains". The Texas Tribune.
13. Cain, C., Lesser, J., & White, B. (2007). A common-sense guide to wholesale electric markets. Online Book, published by Bates White Economic Consulting.
14. Rahimi, A. F., & Sheffrin, A. Y. (2003). Effective market monitoring in deregulated electricity markets. IEEE Transactions on Power systems, 18(2), 486-493.
15. Biswas, S., & Goehring, L. (2019). Load dependence of power outage statistics. Europhysics Letters, 126(4), 44002–.

16. Haifeng S., Zhaoyu W., Jianhui W., Zhen H., Carrington, N., and Jianxin L., (2016). Data-Driven Power Outage Detection by Social Sensors. IEEE Transactions on Smart Grid, 7(5), 2516–2524.
17. Carlsson, F., & Martinsson, P., (2008). Does it matter when a power outage occurs? — A choice experiment study on the willingness to pay to avoid power outages. Energy Economics, 30(3), 1232–1245.
18. Carnero, M.C., & Gomez, A., (2017). Maintenance strategy selection in electric power distribution systems. Energy 129 255 – 272.
19. Flamenbaum R. D., Pompo T., Havenstein C., and Thiemsuwan J., (2019). Machine Learning in Support of Electric Distribution Asset Failure Prediction. SMU Data Science Review: Vol. 2: No. 2, Article 16.
20. Wang, Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters, 136, 190–197. https://doi.org/10.1016/j.patrec.2020.05.035.
21. Warsono, W., Russel, E., Wamiliana, W., Widiarti, W., & Usman, M. (2019). Vector autoregressive with exogenous variable model and its application in modeling and forecasting energy data: Case study of PTBA and HRUM energy. International Journal of Energy Economics and Policy, 9(2), 390-398.
22. Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics, 15(1), 1-9.
23. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.
24. Wang, Y., Sun, S., Chen, X., Zeng, X., Kong, Y., Chen, J., Guo, Y., & Wang, T. (2021). Short-term load forecasting of industrial customers based on SVMD and XGBoost. International Journal of Electrical Power & Energy Systems, 129, 106830
25. Chawla, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. The Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953.
26. PowerOutage.US. Retrieved November 6, 2021, from https://poweroutage.us.
27. One Call API: weather data for any geographical coordinate. Retrieved November 6, 2021, from https://openweathermap.org/api/one-call-api.
28. Hourly Load Data Archives. Retrieved November 6, 2021, from http://www.ercot.com/gridinfo/load/load_hist.
29. Market Prices. Retrieved November 6, 2021, from http://www.ercot.com/mktinfo/prices.
30. Autocorrelation - Wikipedia. Retrieved November 6, 2021, from https://en.wikipedia.org/wiki/Autocorrelation.
31. Hurricane Harvey even summary - energy.gov. (n.d.). Retrieved November 30, 2021, from https://www.energy.gov/sites/prod/files/2017/08/f36/Hurricane%20Harvey%20Event%20Summary%20%231.pdf.
32. Hurricane Harvey - Wikipedia. Retrieved November 29, 2021, from https://en.wikipedia.org/wiki/Hurricane_Harvey.
33. Winter Storm Uri 2021. Retrieved November 29, 2021, from https://comptroller.texas.gov/economy/fiscal-notes/2021/oct/winter-storm-reform.php.
34. Resource Adequacy Challenges in Texas. Retrieved November 29, 2021, from https://www.edf.org/sites/default/files/documents/EDF-ERCOT-Report.pdf.
35. Luo, Zhang, Z., Fu, Y., & Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results in Physics, 27, 104462–104462. https://doi.org/10.1016/j.rinp.2021.104462.

36. Regression in the face of messy outliers? Try Huber regressor, from https://towardsdatascience.com/regression-in-the-face-of-messy-outliers-try-huber-regressor-3a54ddc12516.
37. Ensemble method https://scikit-learn.org/stable/modules/ensemble.html#ensemble.
38. Random forest, from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
39. Gradient boosting, from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html#sklearn.ensemble.GradientBoostingRegressor.
40. Support Vector Machine, https://scikit-learn.org/stable/modules/svm.html#svm.
41. Multi-layer perceptron: https://scikit-learn.org/stable/modules/neural_networks_supervised.html.
42. Pywebio application (n.d) https://www.pyweb.io/.
43. 2021 Texas power crisis. https://en.wikipedia.org/wiki/2021_Texas_power_crisis.
44. Texas Has Nation's Largest Annual State Population Growth. Retrieved February 19, 2022, from https://www.census.gov/library/stories/2017/08/texas-population-trends.html.
45. Texas Water Development Board: Home. Retrieved February 19, 2022, from https://www.twdb.texas.gov/.
46. Understanding Texas' energy grid failure | Harvard Kennedy School. Retrieved February 19, 2022, from https://www.hks.harvard.edu/faculty-research/policy-topics/environment-energy/understanding-texas-energy-grid-failure.

Git Hub: https://github.com/VenkataVanga/MSDS-6120_Capstone.

Appendix:

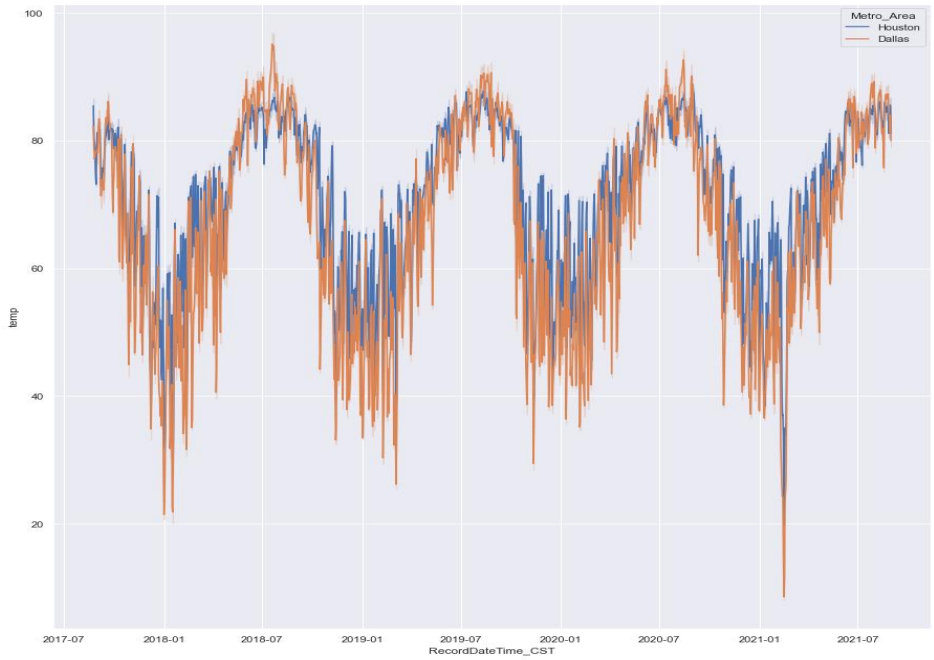
Appendix - Table 1: Data Description.

Item	Variable	Type	Description
1	'Metro_Area'	varchar	The metro area of the county, either Dallas or Houston
2	'State'	varchar	State, will be TX for all entries
3	'County'	varchar	The specific Texas county
4	'Weekday'	Int	Weekday in the month
5	'Month'	Int	Month of the year
6	'hour'	Int	Hour of the day when data is collected
7	'RecordDateTime_CST'	date time	time stamp in CST
8	'RecordDateTime_UTC'	date time	time stamp in UCT
9	'RecordDateTime_EST'	date time	time stamp in EST
10	'Customer_Count'	float	Overall electric Customer Count of the county at that specific time stamp
11	'Outage_Count'	float	Number of customer experiencing an outage at that specific time stamp
12	'temp'	float	temperature degrees in Fahrenheit
13	'feels_like'	float	feels like temperature degrees in Fahrenheit
14	'temp_min'	float	the min temperature degrees in Fahrenheit for the hour ending of the time stamp
15	'temp_max'	float	the max temperature degrees in Fahrenheit for the hour ending of the time stamp
16	'pressure'	float	Atmospheric pressure on the sea level, hPa
17	'humidity'	float	Humidity, %
18	'wind_speed'	float	windspeed in miles/hour.
19	'wind_deg'	float	wind direction in degrees
20	'rain_1h'	float	inches of rain in last hour
21	'rain_3h'	float	inches of rain in last 3 hours
22	'snow_1h'	float	inches of snow in last hour
23	'snow_3h'	float	inches of snow in last 3 hours
24	'clouds_all'	float	Cloudiness, %
25	'ERCOT_WEATHERZONE_LOAD'	float	Total MWH for ERCOT Weather zone for time stamp hour ending
26	'ERCOT_RT_LOADZONE_PRICE'	float	Current Load Zone energy weighted price/MWH for ERCOT load zone hour ending
27	'ERCOT_RT_HUB_PRICE'	float	Current HUB energy weighted price/MWH for ERCOT load zone hour ending
28	'ERCOT_DA_LOADZONE_PRICE'	float	Day Ahead Load Zone energy weighted price/MWH for ERCOT load zone hour ending
29	'ERCOT_DA_HUB_PRICE'	float	Day Ahead HUB energy weighted price/MWH for ERCOT load zone hour ending
30	'Clear_sky is clear'	Boolean	True/False flag for weather parameters and the description
31	'Clouds_broken clouds'	Boolean	True/False flag for weather parameters and the description
32	'Clouds_few clouds'	Boolean	True/False flag for weather parameters and the description
33	'Clouds overcast clouds'	Boolean	True/False flag for weather parameters and the description
34	'Clouds_scattered clouds'	Boolean	True/False flag for weather parameters and the description
35	'Drizzle_drizzle'	Boolean	True/False flag for weather parameters and the description
36	'Drizzle_heavy intensity drizzle'	Boolean	True/False flag for weather parameters and the description
37	'Drizzle_light intensity drizzle'	Boolean	True/False flag for weather parameters and the description
38	'Dust_dust'	Boolean	True/False flag for weather parameters and the description
39	'Fog_fog'	Boolean	True/False flag for weather parameters and the description
40	'Haze_haze'	Boolean	True/False flag for weather parameters and the description
41	'Mist_mist'	Boolean	True/False flag for weather parameters and the description
42	'Rain_extreme rain'	Boolean	True/False flag for weather parameters and the description
43	'Rain_freezing rain'	Boolean	True/False flag for weather parameters and the description
44	'Rain_heavy intensity rain'	Boolean	True/False flag for weather parameters and the description
45	'Rain_heavy intensity shower rain'	Boolean	True/False flag for weather parameters and the description
46	'Rain_light rain'	Boolean	True/False flag for weather parameters and the description
47	'Rain_moderate rain'	Boolean	True/False flag for weather parameters and the description
48	'Rain_proximity shower rain'	Boolean	True/False flag for weather parameters and the description
49	'Rain_shower rain'	Boolean	True/False flag for weather parameters and the description
50	'Rain_very heavy rain'	Boolean	True/False flag for weather parameters and the description
51	'Smoke_smoke'	Boolean	True/False flag for weather parameters and the description
52	'Snow_heavy snow'	Boolean	True/False flag for weather parameters and the description
53	'Snow_light rain and snow'	Boolean	True/False flag for weather parameters and the description
54	'Snow_light snow'	Boolean	True/False flag for weather parameters and the description
55	'Snow_snow'	Boolean	True/False flag for weather parameters and the description
56	'Squall_squalls'	Boolean	True/False flag for weather parameters and the description
57	'Thunderstorm_proximity thunderstorm'	Boolean	True/False flag for weather parameters and the description
58	'Thunderstorm_proximity thunderstorm with rain'	Boolean	True/False flag for weather parameters and the description
59	'Thunderstorm_ragged thunderstorm'	Boolean	True/False flag for weather parameters and the description
60	'Thunderstorm_thunderstorm'	Boolean	True/False flag for weather parameters and the description
61	'Thunderstorm_thunderstorm with heavy rain'	Boolean	True/False flag for weather parameters and the description
62	'Thunderstorm_thunderstorm with light rain'	Boolean	True/False flag for weather parameters and the description
63	'Thunderstorm_thunderstorm with rain'	Boolean	True/False flag for weather parameters and the description

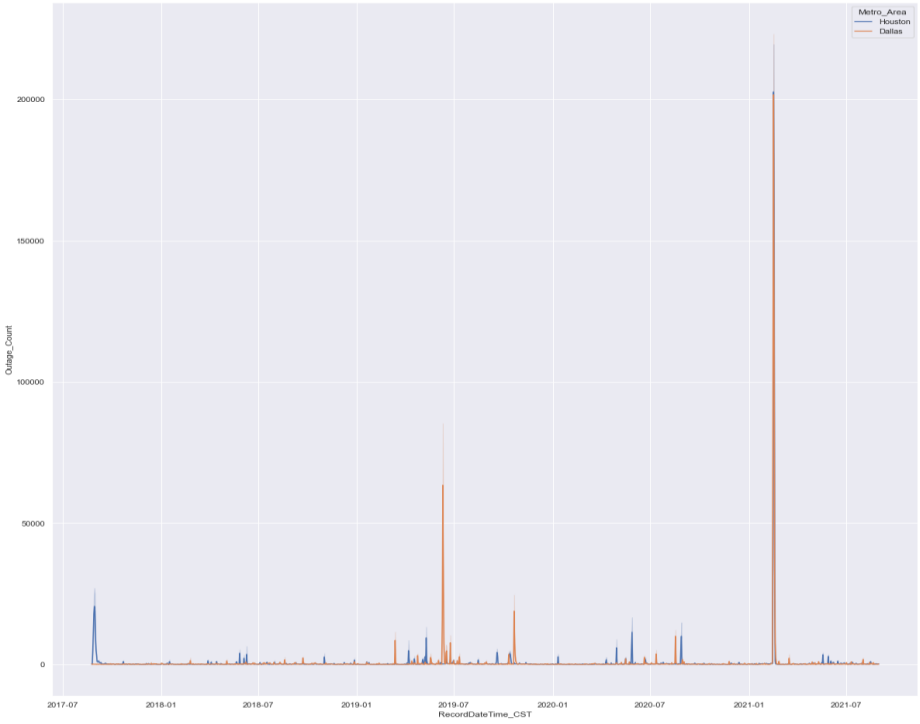
	datetime	Metro_Area	RecordDateTime_CST	RecordDateTime_UTC	State	County	RecordDateTime_EST	Customer_Count	Outage_Count	temp
0	2021-11-03 17:30:48.913	Houston	2018-09-17 17:00:00	2018-09-17 22:00:00	Texas	Brazoria	2018-09-17 18:00:00	122933	13	91.74
1	2021-11-03 17:30:48.913	Houston	2017-08-27 07:00:00	2017-08-27 12:00:00	Texas	Brazoria	2017-08-27 08:00:00	105618	2349	77.76
2	2021-11-03 17:30:48.913	Houston	2019-11-06 08:00:00	2019-11-06 14:00:00	Texas	Brazoria	2019-11-06 09:00:00	123921	11	66.94
3	2021-11-03 17:30:48.913	Houston	2018-09-20 14:00:00	2018-09-20 19:00:00	Texas	Brazoria	2018-09-20 15:00:00	122933	66	88.70
4	2021-11-03 17:30:48.913	Houston	2018-09-21 10:00:00	2018-09-21 15:00:00	Texas	Brazoria	2018-09-21 11:00:00	122933	3	80.80
...	...	...	...	...	...	...	...	...	...	...
317019	2021-11-03 17:30:48.913	Dallas	2018-01-16 12:00:00	2018-01-16 18:00:00	Texas	Tarrant	2018-01-16 13:00:00	781583	19	24.84
317020	2021-11-03 17:30:48.913	Dallas	2019-03-28 21:00:00	2019-03-29 02:00:00	Texas	Tarrant	2019-03-28 22:00:00	860771	10	71.49
317021	2021-11-03 17:30:48.913	Dallas	2019-03-30 01:00:00	2019-03-30 06:00:00	Texas	Tarrant	2019-03-30 02:00:00	860766	15	67.50
317022	2021-11-03 17:30:48.913	Dallas	2019-03-30 16:00:00	2019-03-30 21:00:00	Texas	Tarrant	2019-03-30 17:00:00	860608	161	53.22
317023	2021-11-03 17:30:48.913	Houston	2018-06-17 23:00:00	2018-06-18 04:00:00	Texas	Montgomery	2018-06-18 00:00:00	241662	32	77.13

317024 rows × 63 columns

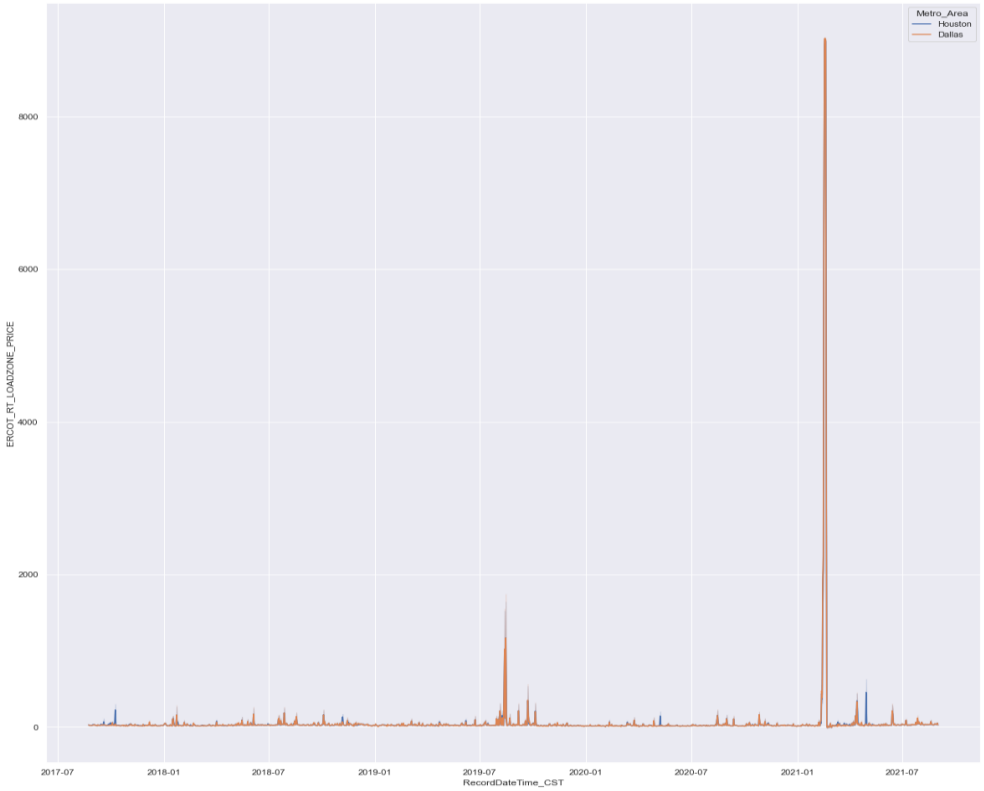
Appendix - Figure B: Data Table



Appendix - Figure C: Temperatures from July 2017 to September 2021 in Dallas & Houston



Appendix - Figure D: Energy outage counts from July 2017 to September 2021 in Dallas & Houston

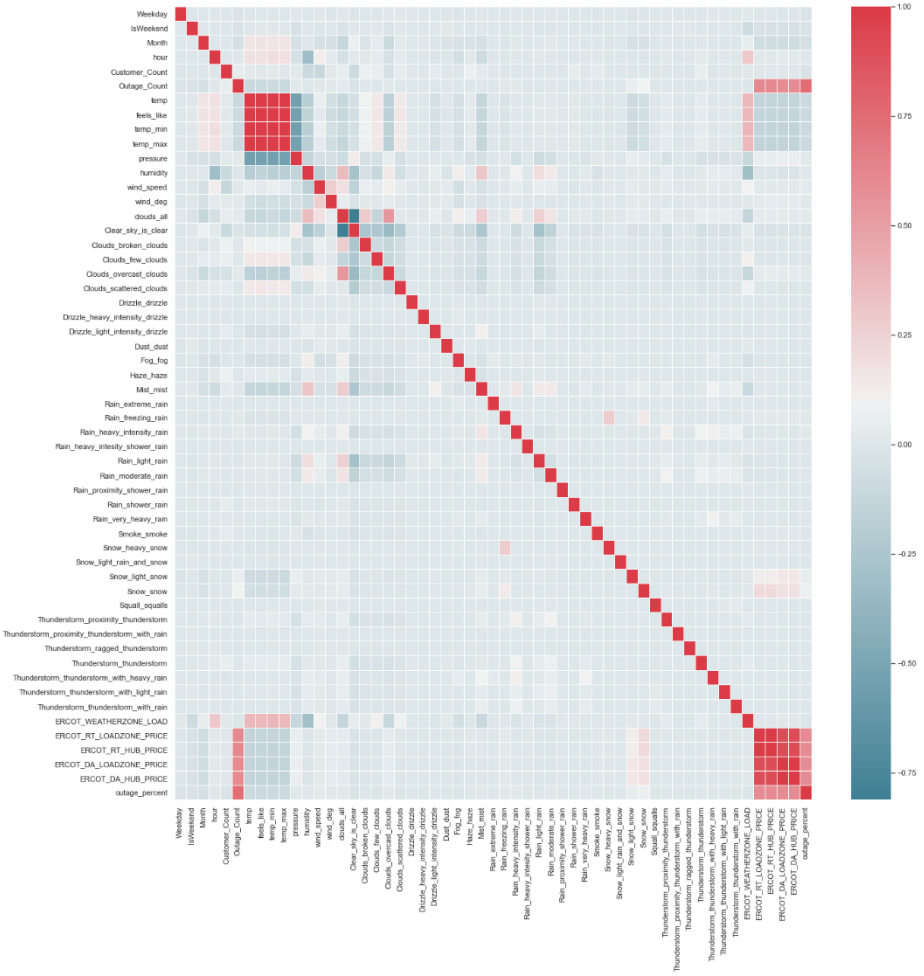


Appendix - Figure E: ERCOT Electricity price of current load zone from July 2017 to September 2021 in Dallas & Houston

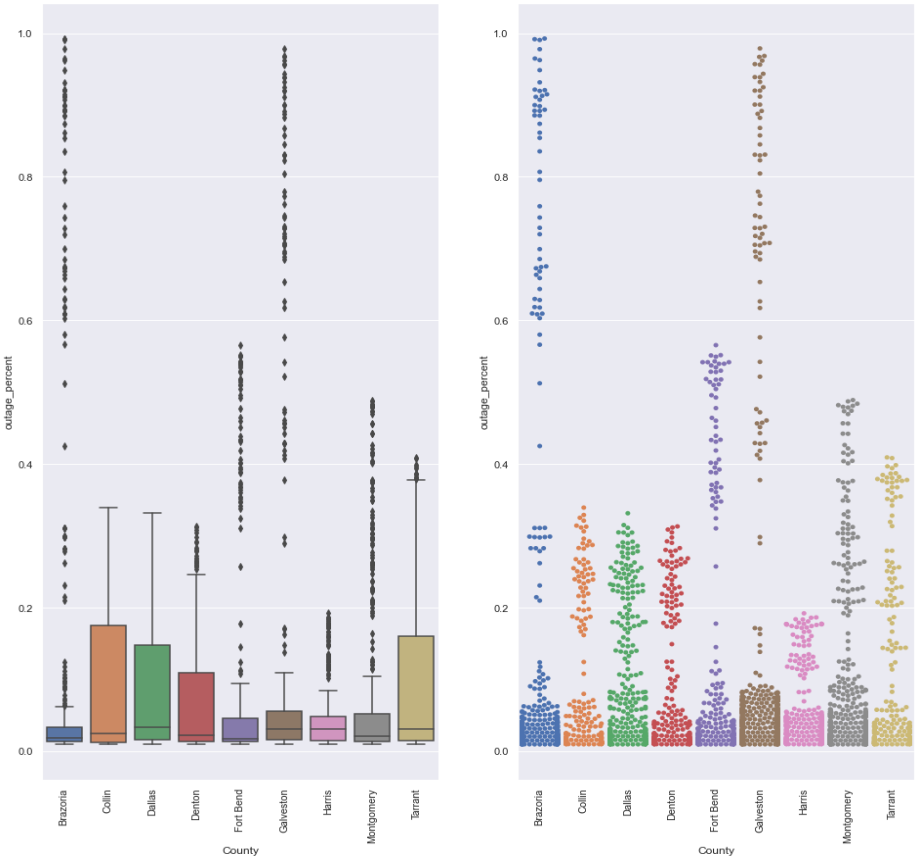
		Outage_Count
Metro_Area	Snow_snow	
Dallas	0	606.8
	1	32878.9
Houston	0	596.0
	1	31599.5

		Outage_Count
Metro_Area	Rain_freezing_rain	
Dallas	0	628.8
	1	26.4
Houston	0	611.7
	1	77524.0

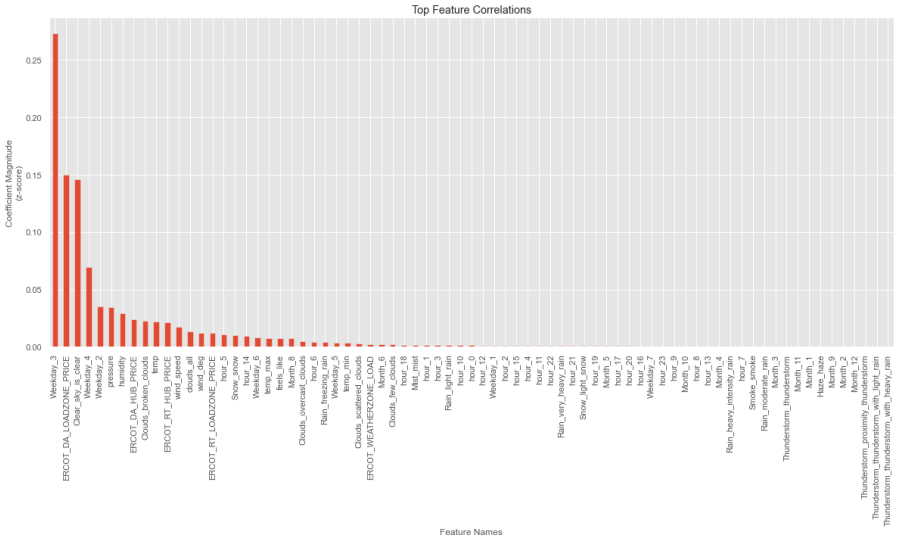
Appendix - Figure F: Outage average counts based on snow and freezing rain from July 2017 to September 2021 in Dallas & Houston



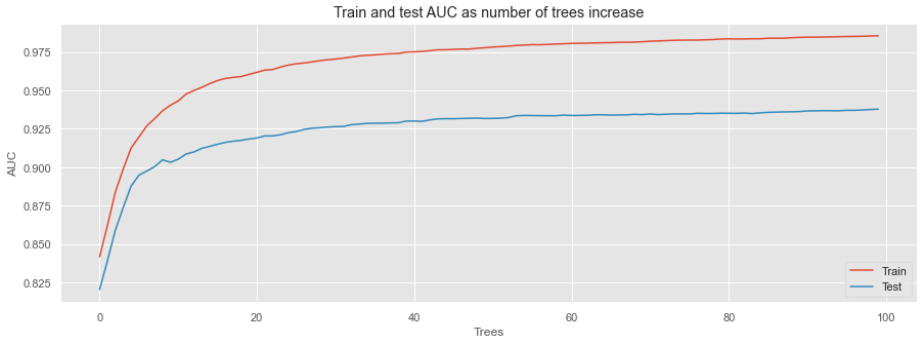
Appendix - Figure G: Correlation heatmap for all continuous variables in the dataset



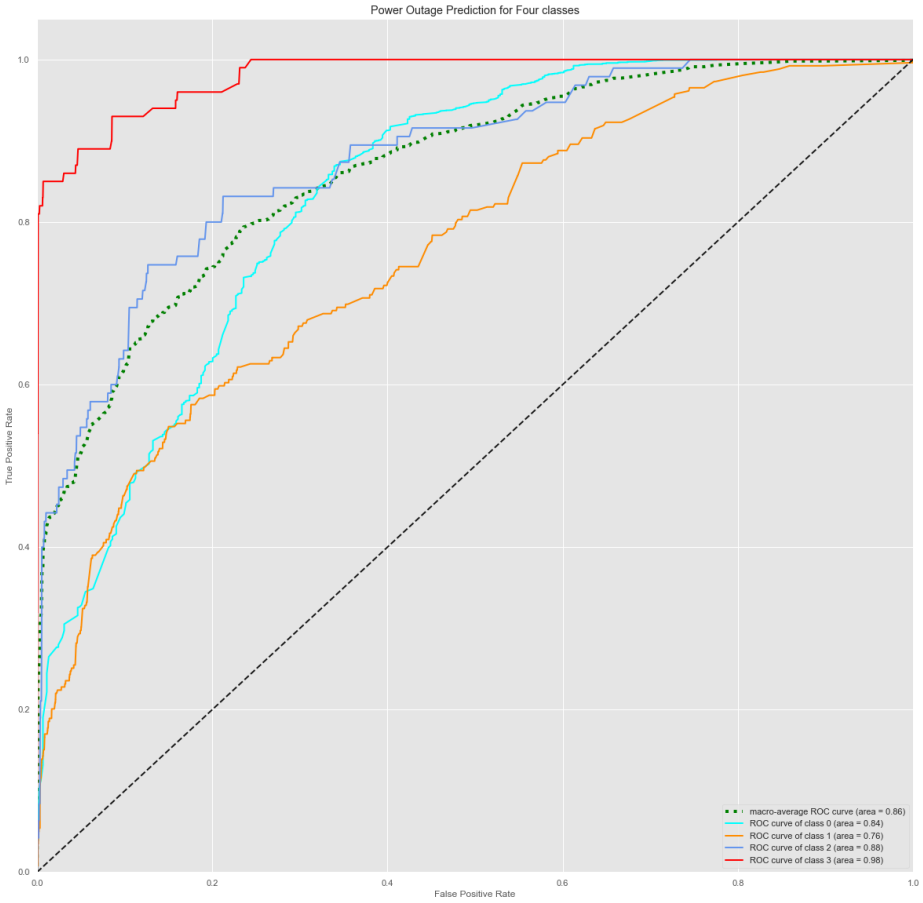
Appendix - Figure H: Outage Percentage boxplot and swam plot with respect to County



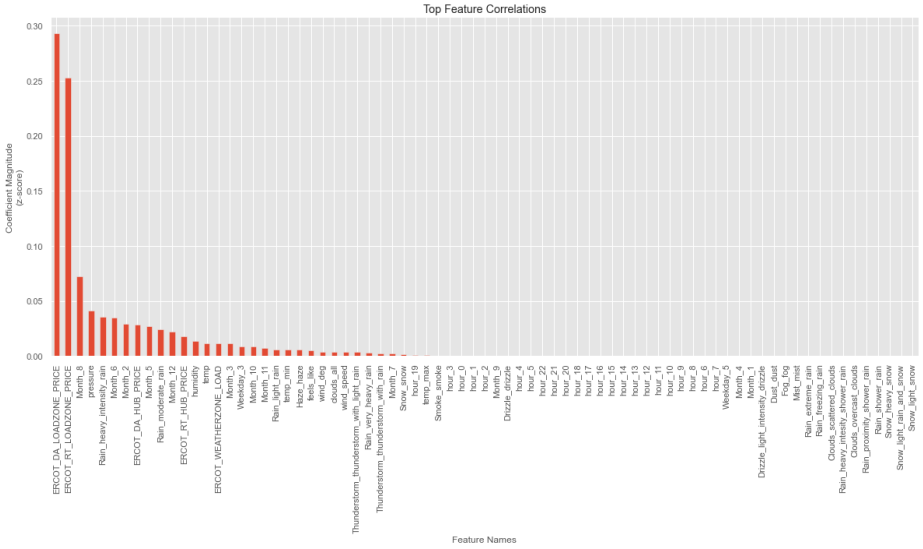
Appendix - Figure I: Feature importance plot from XGBoost Regression



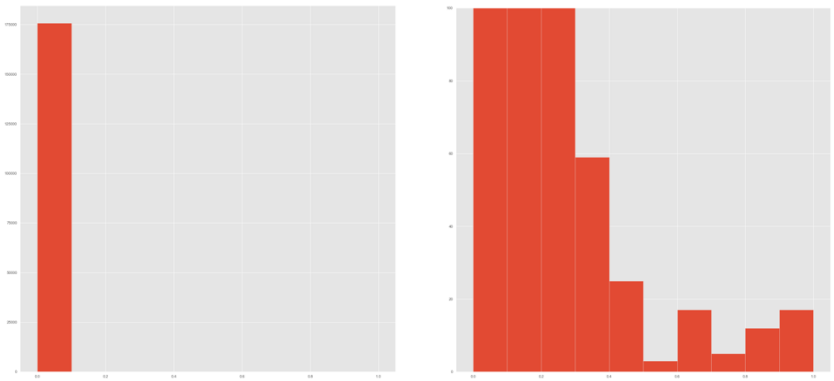
Appendix - Figure J: Train and test ROC plot of XGBoost Classification



Appendix - Figure K: ROC plot of XGBoost Classification for all classes



Appendix - Figure L: Feature importance plot from XGBoost Classification



Appendix - Figure M: Response variable distribution