CSCE 5222: Feature Engineering

Project Proposal

Project Title: "Power Outage Insight: Feature Engineering in Time Series Analysis"

Team Members:

1. Sravani Katlaganti (11560617)

2. Panduga Raja Tejasvi Prasad (11414926)

3. Yasmeen Haleem (11462753)

4. Varun Mohan (11615500)

GitHub Link: https://github.com/yasmeenha/Featureengg

Idea Description:

As time series feature engineering deals with lags and temporal and spatial analysis, the team decided to work on the Eagle-I power dataset from the SMC dataset challenge 2023, and after studying the ieee paper[5] the team decided to apply feature engineering techniques like lags, autocorrelation, dickey fuller test to find anomalies and seasonal decomposition, and the final propose is to see the improvement in power outage prediction model using LSTM, SVM, linear regression, xgboost and cat boost regressor.

Goal and Objectives:

As the power outage dataset from Eagle-I only has the given outages as sum as the column with the power outages of 92% of U.S county with 15-minute time stamp, the team decided to do feature engineering on time series Eagle-I dataset, and to do temporal and spatial analysis on the power outage data. The objective is to combine the given power outage data from 2014 -2022 in the Eagle-I dataset and to combine it with the U.S weather dataset from Kaggle that gives the weather information for U.S counties from 2016 to 2022, such as rain, fog, snow, wind speed, etc. One aim is to find correlation between weather and power outages and after doing time series feature engineering the team aims to see an improvement in the accuracy of the power outage prediction. The other aim is to create a multivariate power outage prediction model by creating rainfall, windspeed, humidity, different days of the week, months as features and try to see if the power outage prediction accuracy improves in a multivariate model.

Motivation:

Power outages are a common and bothersome concern which impacts individuals, firms, and critical infrastructure. The need to address the issues that are due to power outages and ultimately improve the dependability of the electric system proved to be the highlight of the project. These outages may be caused by a variety of reasons, such as hardware malfunctions, severe weather, natural disasters, and even cyberattacks. We have to comprehend these factors and work towards reducing the incidences and duration of power outages by enhancing methods of feature engineering for time series data pertaining to power outages [4].

By performing feature engineering on time series and taking into account the effect of weather correlation on the power outages we try to see the improvement in the power outage prediction and offer utility companies, enterprises, and governments useful information to help them make decisions, prioritize grid enhancements, and allocate resources more effectively.

Significance:

The significance of this project is diverse, involving a number of important elements that have an impact on both individuals and society as a whole. Firstly, this study has the potential to significantly improve the precision and efficacy of predictive analytics by concentrating on feature engineering for time series data related to power outages. These more accurate forecasts can result in earlier alerts and more effective preventive measures, which will eventually reduce the incidence and length of power outages. By decreasing downtime, minimizing equipment damage, and raising overall output, this increase in grid stability and resilience offers a superior quality of life for people and businesses [2].

The financial consequences are additionally crucial. Due to business being disrupted and the requirement for backup power sources, power outages frequently result in large financial losses. By using advanced feature engineering to estimate the economic cost of power outages, this study can offer utilities and companies useful information to help them make informed choices about grid upgrades, resource allocation, and investment. In turn, this results in significant cost savings and more effective resource usage, which is advantageous to the economy and consumers.

Furthermore, it impacts healthcare facilities, communication networks, and emergency response procedures. The project directly contributes to public safety and the efficient operation of vital services during time-sensitive situations by improving grid dependability through improved feature engineering. The relevance for crucial services by making timely and accurate predictions that can make difference between life and death in healthcare situations and prompt reactions to emergencies [3].

Literature survey:

The study of time series data and feature engineering come together crucially in the predictive modeling of power outages, highlighting the importance of managing temporal data skillfully and creating features that reflect the inherent dynamics and patterns in the data. In this regard, a thorough examination of numerous feature extraction and selection techniques as well as forecasting methodologies offers a fundamental comprehension of the dominant tactics and difficulties in this field.

Sandya H.B. et al. illuminate the importance of intelligent feature extraction, specifically employing Fuzzy Logic and GARCH techniques in time-series signal processing [1].Although the methodologies presented may not directly relate to the prediction of power outages, they do highlight a key idea in feature engineering: the extraction of pertinent features that accurately capture the temporal dependencies and volatilities present in electrical grid data. These features may be useful in improving predictive models for power outages.

Khalid Ijaz et al. present a novel approach to feature selection in the temporal domain, proposing an LSTM model integrated with a distinctive temporal feature selection technique for short-term electrical load forecasting [2]. Particularly in the field of power outage prediction, where temporal dependencies are critical, the emphasis on choosing salient temporal features that can accurately describe the sequential data highlights the pivotal importance of feature selection in increasing model performance and generalization.

Vivian Do et al. delve into the spatiotemporal aspects of power outages, exploring the intricate relationships with climate events and social vulnerability [3]. This research highlights the possibilities of integrating temporal elements, together with social and climatic factors, into predictive models in order to help them recognize and respond to anomalies and patterns connected to weather conditions and social interactions. By offering a multidimensional approach that incorporates temporal, geographical, and external elements, this broadens the scope of feature engineering.

Finally, a comparison of feature extraction and selection methods for time-series data classification can shed light on the advantages and disadvantages of various approaches, laying the groundwork for effective feature engineering approaches for power outage prediction [4]. Exploration and adaptation of these approaches may open the door to more sophisticated and effective models that can recognize and respond to trends and abnormalities in data from the electrical grid.

Features:

Dataset: The dataset that we are going to use in this project is the Eagle-I power dataset from the SMC dataset challenge 2023[6]. The dataset consists of 8 years of power outage data between 2014 and 2022 for each county of united states with an interval of 15 mins which is derived by the EAGLE-I program at ORNL. They have collected it through a process of ETL from public outage maps of utility. The columns present in this dataset are 'fips_code', 'county', 'state', and 'sum' indicating the power outage and the total number of rows are 1689460. The other dataset that we will use to integrate with the former dataset is the weather dataset from Kaggle dataset of US Weather Events (2016 - 2022)[7]. This dataset consists of 7 years of weather events data between January 2016 and December 2022 for each county of 49 states in the US which is collected from 2071 airport-based weather stations nationwide. It consists of events of about 8.6 million including general rain and snow to severe storms and freezing conditions. The columns include 'ZipCode', 'County', 'AirportCode', 'EventId'. 'Type', 'Severity' etc. We'll be integrating these 2 datasets for further use in the feature engineering and modeling processes.

Time Series Feature Engineering: The above integrated dataset is then used for feature engineering methods of time series. Some of these methods are to apply lag, autocorrelation and differencing that will produce new and improved features for better understanding and better performance of the models. The resulting features will also include some special and temporal information and the target feature dependencies.

Seasonal Decomposition: This project also involves seasonal decomposition methods as our data is time series data. These decomposition methods are applied to get differentiated components of seasonality, trends and residues. These are useful for better understanding the patterns and variations present in the data.

Data Integration and Correlation Analysis: We plan to integrate 2 datasets that are discussed above namely Eagle-I power dataset and US Weather dataset based on county column. The resulting dataset containing weather data and power outage data is used to analyze the correlation between weather events and power outages. This analysis involves performing tests of statistics and data visualization that results in the explanation of how power outage is dependent on weather conditions.

Multivariate ML Prediction Models: As the input data consists of multiple features including rainfall, humidity, wind speed, days and months to forecast power outages, we are planning to develop a multivariate predictive model with algorithms of LSTM, SVM, linear regression, xgboost and cat boost regressor. Then train the models using training data and test the model using testing data.

Evaluation: The above models are then evaluated. Using these improved features and weather data will result in increased accuracies and better performance of the models when compared to the models without these engineered features.

Expected Outcome:

The main expected outcome of this project is to increase the performance of the machine learning models that are used for forecasting the power outage. The ultimate aim of the team is to build more accurate predictive models with the application of feature engineering methods to and also by combining the weather data to produce more improved and informative features for the better prediction of power outage.

References:

[1] https://ieeexplore.ieee.org/document/6851555
[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9849665
[3] https://www.nature.com/articles/s41467-023-38084-6
[4] https://ieeexplore.ieee.org/abstract/document/9763125
[5] https://www.osti.gov/biblio/1430039
[6] https://smc-datachallenge.ornl.gov/eagle/
[7] https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events/data