Power Outage Insights: Feature Engineering in Time Series Analysis

1. Sravani Katlaganti (11560617)
2. Panduga Raja Tejasvi Prasad (11414926)
3. Yasmeen Haleem (11462753)
4. Varun Mohan (11615500)

Introduction

- **Data Integration**: Combine power outage and weather data for Lee and Miami counties, Florida (2018-2022).
- **Feature Engineering**: Utilize time-based features, rolling statistics, and lag analysis.
- **Modeling Techniques**: Employ ARIMA, SARIMA, Linear Regression, XGBoost, ARIMAX, CatBoost Regressor, SVM and LSTM models.
- **Performance Metrics**: Assess models using RMSE, MAE, and R^2.
- **Practical Insights**: Aid in improving power outage prediction and utility management.

Phase-1

Module	Description	Action Item	Member	%
1	Data Preprocessing and a model	• Cleaning and merging data • Xgboost model • Coding and Documentation	Varun Mohan	5 5 5 5 5
2	Time Series Feature Engineering and a model	• Time series analysis • SARIMA model • Coding and Documentation	Panduga Raja Tejasvi Prasad	5 5 5 5 5
3	Data Visualization and a model	• Data Visualization • ARIMA model • Coding and Documentation	Yasmeen Haleem	5 5 5 5 5
4	A model and Detailed Analysis of Outage Prediction Models	• Linear Regression model • Result Analysis for models • Coding and Documentation	Sravani Katlaganti	5 5 5 5 5

Phase-2

Module	Description	Action Item	Member	%
1	Analysis for different counties and model prediction	• SVM model • Coding and Documentation	Varun Mohan	25%
2	Other prediction models implementation and analyzing for different county	• Catboost Regression model • Coding and Documentation	Panduga Raja Tejasvi Prasad	25%
3	Modeling for different counties and model prediction	• LSTM Model • Coding and Documentation	Yasmeen Haleem	25%
4	Model prediction and Detailed Analysis of Outage Prediction Models from different models and counties	• ARIMAX model • Coding and Documentation	Sravani Katlaganti	25%

Dataset

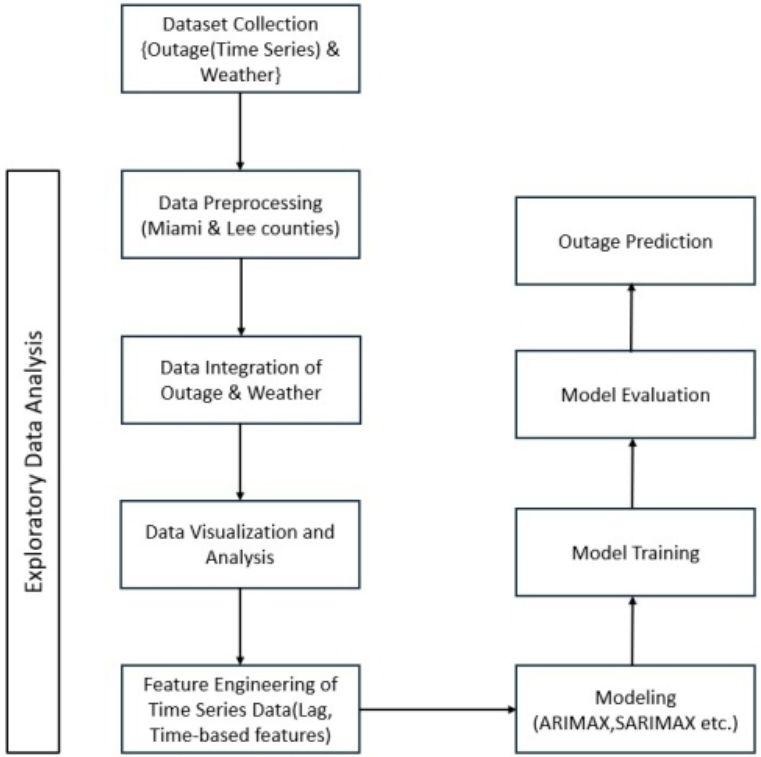
Eagle-I power dataset

- **Dataset Size:** Consists of 1,689,460 rows with 'POWER OUTAGE' data, span over 4 years from 2018 - 2022 for each county of United States with an interval of 15 mins.
- **Columns:**
  - 'fips_code': Federal Information Processing Standards (FIPS) code for unique county identification.
  - 'county': County name.
  - 'state': State of the county.
  - 'sum': Indicates power outage occurrences.

US Weather Events (2016 - 2022)

- **Dataset Size:** Total events of 8.6 million. The dataset contains information about various weather events from January 2016 to December 2022.
- **Columns:**
  - Event type: Type of the weather event (e.g., Fog, Rain).
  - Severity: Severity level of the weather event (e.g., Severe, Light).
  - Starttime:Timestamp indicating the start time of the weather event
  - Endtime :Timestamp indicating the end time of the weather event
  - County:County where the event occurred.
  - State:State where the event occurred.

Workflow


Dataset Collection {Outage(Time Series) & Weather}
Data Preprocessing (Miami & Lee counties)
Data Integration of Outage & Weather
Data Visualization and Analysis
Feature Engineering of Time Series Data(Lag, Time-based features)
Modeling (ARIMAX,SARIMAX etc.)
Model Training
Model Evaluation
Outage Prediction
Exploratory Data Analysis

Exploratory Data Analysis

Data Preprocessing of Outage Dataset:

- Filtering Data
- Date-Time Conversion and Manipulation
- Grouping and Aggregation

Data Preprocessing of Weather Dataset:

- Column Dropping
- Date-Time Conversion and Localization:
- Date Filtering(Drop rows for the years 2016 and 2017)
- Severity Mapping
- Duplicate Removal

Data Integration:

- Merge the two DataFrames on their 'date'/'Date' columns using an outer join.
- Handling missing values

Time Series Feature Engineering

- **Time-based features (**Time based features such as Year, Month and Day)
- **Distribution Analysis** (Rain, Fog, Cold, Precipitation, Storm, Hail)
- **Lag Features**
- **Rolling Windows**
- **Date-Time Decomposition**

Power Outage Prediction Models

- **ARIMA (AutoRegressive Integrated Moving Average)**
- **SARIMAX (Seasonal ARIMA)**
- **Linear Regression**
- **XGBoost (eXtreme Gradient Boosting)**
- **ARIMAX**
- **Catboost Regressor**
- **SVM**
- **LSTM**

1. ARIMA (AutoRegressive Integrated Moving Average)

The important steps included checking for stationarity using dickey fuller test, determining ARIMA parameters (p, d, q), and fitting the model to the data.It was observed that ARIMA models are well-suited for time series data without external influences.

2. LSTM

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) specifically designed to learn long-term dependencies in sequence data, which makes them highly effective for predicting power outages. They can analyze complex sequences and temporal patterns in historical power usage, weather conditions, and other relevant data to detect anomalies or conditions likely to lead to a power outage, thereby aiding in prevention and timely response.

3. XGBoost (eXtreme Gradient Boosting)

XGBoost is implementation of gradient boosting algorithms in advanced way, and is known for its efficiency, flexibility, and portability and it makes multiple decision trees sequentially, where each tree tries to correct the errors of its predecessor. This model can handle various types of data, and non-linear relationships and interactions between variables. It usually offers high predictive accuracy and is efficient on large datasets.

4. SVM

SVM regression or Support Vector Regression (SVR) is a machine learning algorithm used for regression analysis. It is different from traditional linear regression methods as it finds a hyperplane that best fits the data points in a continuous space, instead of fitting a line to the data points.

5. Catboost Regressor

CatBoost is a supervised machine learning method that is used by the Train Using AutoML tool and uses decision trees for classification and regression. As its name suggests, CatBoost has two main features, it works with categorical data (the Cat), and it uses gradient boosting (the Boost)

6. SARIMAX (SARIMA with eXogenous variables)

Combined the seasonal features of SARIMA with the ability to incorporate exogenous variables. It required careful selection of both seasonal and non-seasonal parameters, along with relevant external predictors.

7. ARIMAX

ARIMAX is an advanced statistical time series model that predicts future points by accounting for the past values of the variable, non-stationarity in the data, past error terms, and the influence of external or independent variables (exogenous factors) on the variable being forecasted, such as predicting power outages by including weather conditions as predictors.

8. Linear Regression

Linear Regression is statistical method used to model the relationship between a dependent variable (power outage sum) and one or more independent variables (like weather conditions).This model assumes a linear relationship between the input variables and the target.

CODE DEMO

Results

Models performance metrics:

County: Lee, Florida

Model	RMSE	MAE	R^2
ARIMA	1366.84	684.56	-0.00
SARIMAX	1368.61	697.59	-0.01
Linear Regression	1371.78	684.92	-0.01
XGBoost	1930.88	834.58	-1.01
ARIMAX	1371.78	684.93	-0.01
CatBoost Regressor	1639.76	768.09	-0.45
SVM	1414.74	393.9	-0.08
LSTM	1442.05	688.48	-0.03

Results

County: Miami-Dade, Florida

Model	RMSE	MAE	R^2
ARIMA	1699.93	1230.64	-0.01
SARIMAX	1697.18	1214.14	-0.0
Linear Regression	1703.14	1225.73	-0.01
XGBoost	1879.99	813.0	-0.23
ARIMAX	1703.34	1226.08	-0.01
Catboost Regressor	1946.44	801.32	-0.32
SVM	1861.92	935.47	-0.21
LSTM	2305.32	1106.94	-0.01

Outage Prediction Model's Performance Analysis

The performance metrics used are:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-Squared (R^2 or the coefficient of determination)

- ARIMA and SARIMA balanced accuracy effectively; Linear Regression was comparable, while XGBoost and LSTM underperformed. The CatBoost Regressor showed potential overfitting with a negative R^2, and SVM had the best MAE, indicating accurate outage predictions.

Conclusion

- Performed Data preprocessing, Exploratory Data Analysis and Feature Engineering.
- Analyzed trends and correlation between the Outage and Weather datasets.
- Models used are ARIMA, SARIMA, Linear Regressor, XGBoost, ARIMAX, Catboost Regressor, SVM, and LSTM.
- Overall, ARIMA is the best performed model.

Thank you