# Electric Outage Duration Prediction - Final Report

## 1. Introduction

This report provides a comprehensive overview of the electric outage duration prediction project. The objective was to develop a predictive model capable of estimating the duration of electric outages based on various features such as the event description, year, and respondent information. Multiple machine learning models were evaluated to determine the best-performing model.

## 2. Data Preprocessing

The dataset used for this project contained information on various electric outages, including features such as 'Event Description,' 'Year,' and 'Respondent.' During preprocessing, categorical variables such as 'Event Description' and 'Respondent' were encoded using label encoding to make them suitable for machine learning models. Additionally, rows containing missing values in the target variable 'Duration (Hours)' were removed.

## 3. Model Training and Evaluation

Several machine learning models were trained and evaluated, including Linear Regression, Decision Tree Regressor, Support Vector Regressor, K-Neighbors Regressor, Gradient Boosting Regressor, Neural Network Regressor, and XGBoost Regressor. The models were evaluated based on their Root Mean Squared Error (RMSE) performance metric on the test dataset. Due to computational constraints, the final evaluation was focused on three models: Linear Regression, Decision Tree Regressor, and K-Neighbors Regressor.

The RMSE values for the evaluated models were as follows:

- Linear Regression: RMSE = 95.65

- Decision Tree Regressor: RMSE = 90.3

- K-Neighbors Regressor: RMSE = 99.63

## 4. Final Model

Based on the evaluation results, the Decision Tree Regressor was selected as the final model due to its superior performance in terms of RMSE. The following parameters were used for the Decision Tree Regressor:

- Criterion: 'squared\_error' (default)

- Splitter: 'best' (default)

- Maximum Depth: None (default)

- Minimum Samples Split: 2 (default)

- Minimum Samples Leaf: 1 (default)

The final RMSE of the Decision Tree Regressor on the test set was 90.3.

## 5. Conclusions and Recommendations

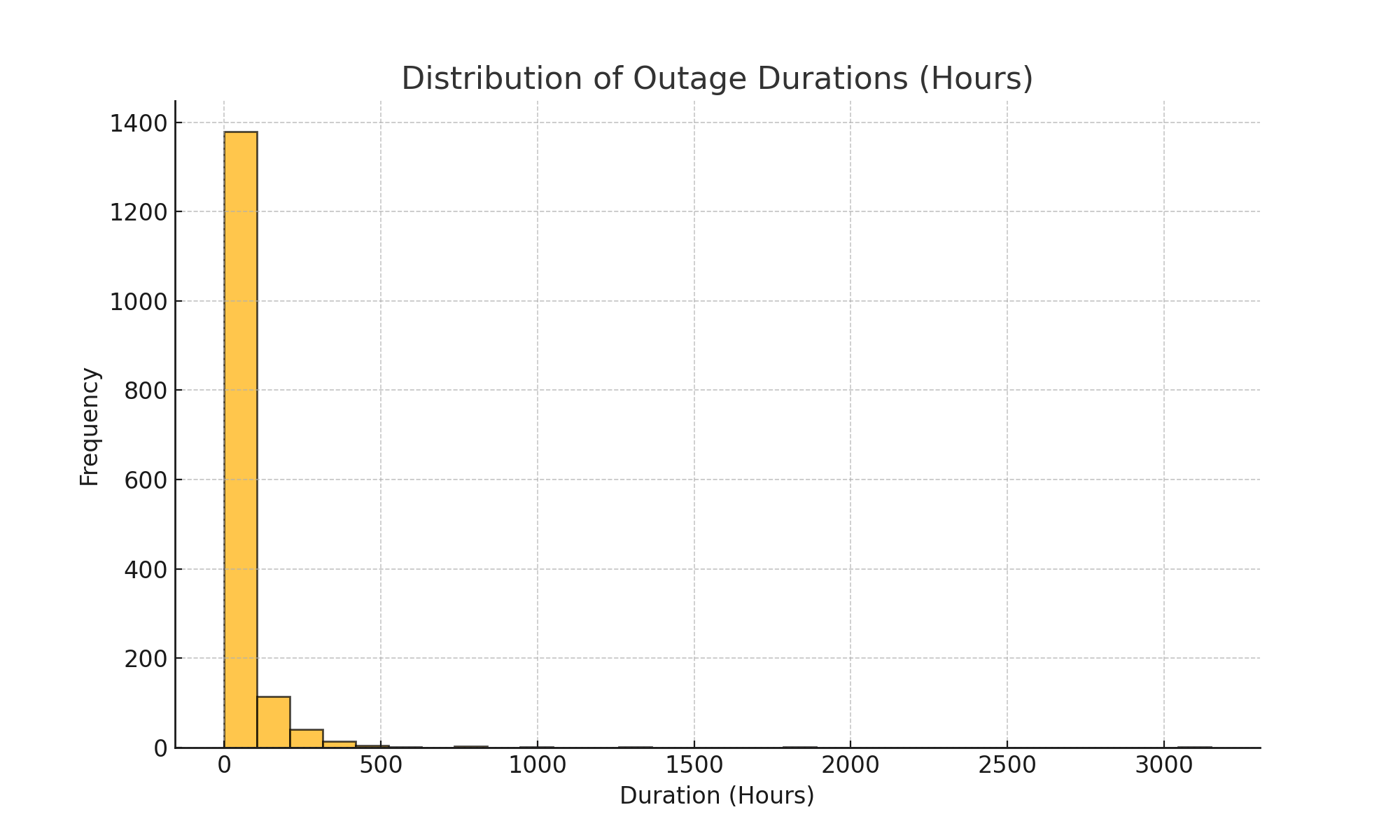
The Decision Tree Regressor provided the most accurate predictions for electric outage durations among the evaluated models. While the model performed reasonably well, there is room for improvement. Future work could involve tuning the hyperparameters of the Decision Tree Regressor or experimenting with ensemble methods such as Random Forest or Gradient Boosting for potentially better accuracy. Additionally, collecting more data and incorporating additional relevant features could further enhance model performance.

# 6. Exploratory Data Analysis (EDA)

This section presents the findings from the Exploratory Data Analysis (EDA) conducted on the electric outage dataset. The purpose of the EDA was to understand the characteristics of the data, identify patterns, and gain insights into the distribution of electric outages and their impact on customers.

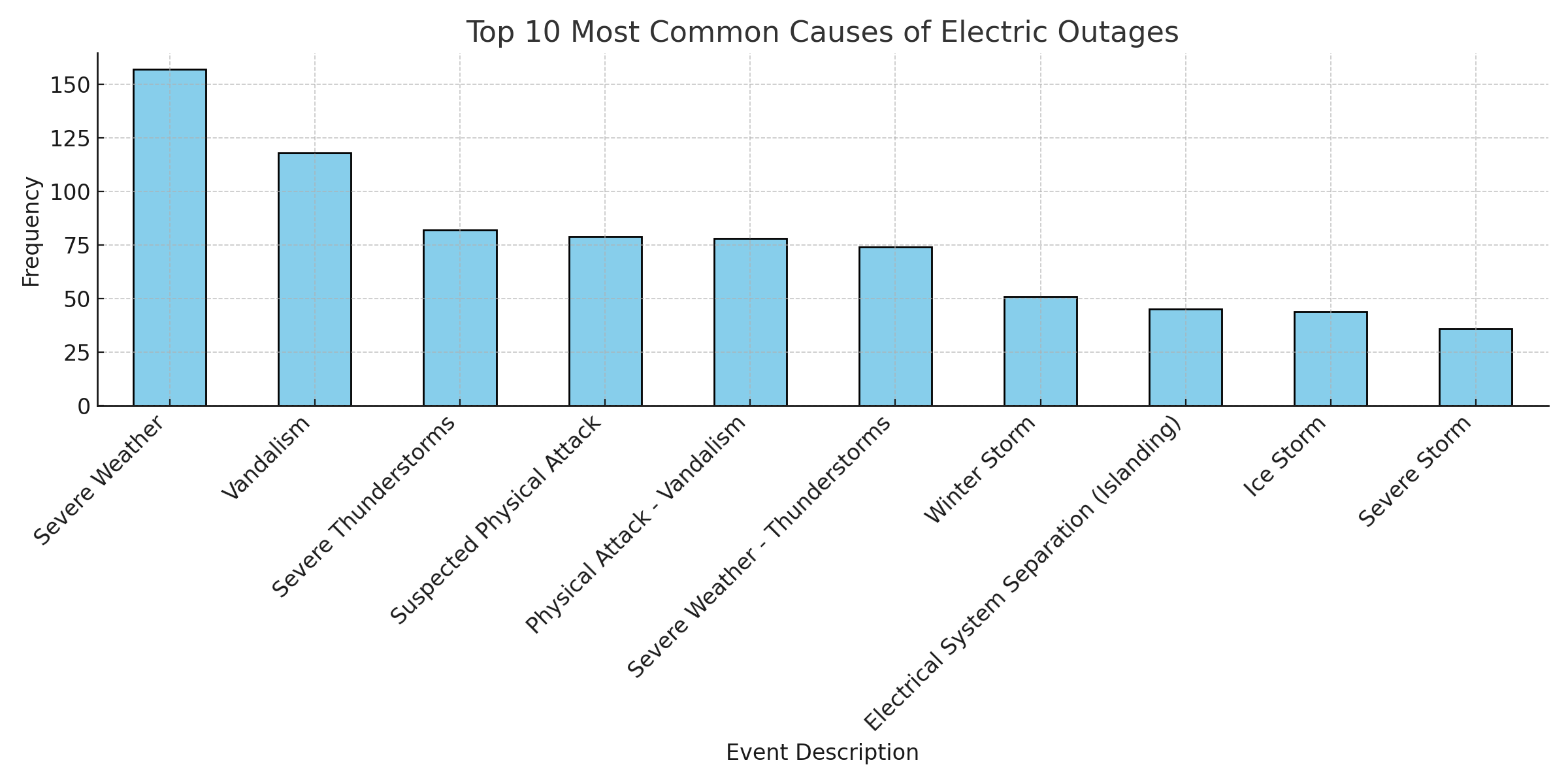
## 6.1 Distribution of Outage Durations

The distribution of outage durations reveals that most electric outages are relatively short, with a small number of events lasting much longer. The histogram below illustrates the distribution of outage durations in hours.



## 6.2 Most Common Causes of Electric Outages

The analysis identified the top 10 most common causes of electric outages. Severe weather events emerged as the leading cause, followed by other factors such as vandalism. The bar chart below provides an overview of the most common causes.



## 6.3 Impact of Outages on Customers

The relationship between outage duration and the number of customers affected shows that longer outages can impact a substantial number of people. The scatter plot below demonstrates this relationship.

