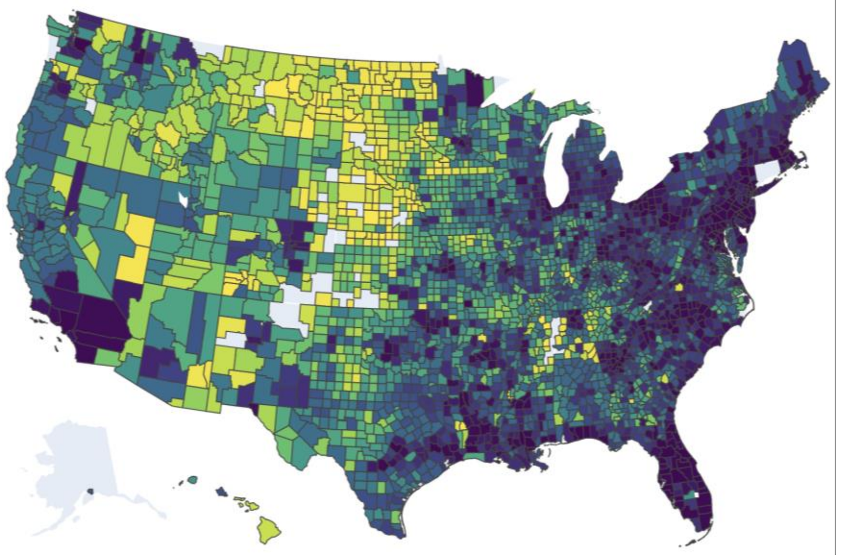

Team:
Ievgen Tsygankov, Liezl Magpayo, Sajid Kolliyath
(Machine Learning Course – University of Toronto SCS)


# Predicting Outages

**Modeling power outages caused by extreme weather**

# Problem

Develop a supervised model to predict power outages and how they correlate with extreme, rare weather events in the USA



- Power outages during extreme weather events cause significant disruptions
- Need for proactive prediction systems to mitigate impact
- **Objective:** Build a robust, data-driven model that can predict outage and its impact based on historical storm and outage data
- Aim is to predict the following:
  - ➢ Occurrence of Power Outages
  - ➢ Duration of outages (in minutes)

Ref: Competition: Dynamic Rhythms thinkonward

# Methodology

- Gather Data – Collected raw data from multiple sources.
- Master Dataset Creation – Built unified, clean dataset for analysis.
- Explore Data – Identified patterns and distributions
- Preprocess Data – Cleaned, normalized,managed outliers, and handle missing values.
- Feature Engineering – Created meaningful inputs from available features
- Shortlist Models – Chose best algorithms for problem type.
- Fine Tune System – Optimized performance through hyperparameter tuning.

**Problem 1: Predict if an outage occurs**

- Binary classification model to detect likelihood of outage based on input data

**Problem 2: Predict how bad an outage will be**

- Regression model estimates outage duration to support planning and mitigation

# Data Sources

### Storm Events Data

- **Source:** NOAA NCEI – Storm Events Database 2014–2024
- **Description:** Comprehensive records of extreme weather events in the U.S.,including event types (e.g., thunderstorms,hurricanes), locations, durations etc
- **Purpose:** Helps identify and quantify the frequency and severity of storms that may cause power outages

### Power Outage Data

- **Source:** EAGLE-I (The Environment for Analysis of Geo-Located Energy Information) 2014–2023
- **Description:** Detailed reports of electricity outages, including time of occurrence, duration, number of customers affected and geographic location
- **Purpose:** Used as the target variable for modeling outage occurrence and severity and understanding historical outage trends

### Tree Canopy Cover

- **Source:** NLCD 2021 – U.S. Forest Service
- **Description:** High-resolution spatial data converted to tabular form using QGIS app to get percentage of land covered by tree canopy
- **Purpose:** Tree cover contributes significantly to storm-related infrastructure damage, especially from falling trees or branches

# Data Sources (contd.)

### Transmission Lines

- **Source:** Homeland Infrastructure Foundation-Level Data (HIFLD)
- **Description:** Spatial data on electrical transmission infrastructure, including the type – underground / overground and route of transmission lines
- **Purpose:** Important for assessing infrastructure exposure and susceptibility to storm impacts

### Meteorological Data

- **Source:** ERA5 – Copernicus Climate Data Store
- **Description:** Daily historical weather data including wind speed, precipitation, temperature, humidity, and pressure
- **Purpose:** Used to enrich the storm dataset and improve model accuracy by capturing detailed environmental conditions during outages

### Socio Economic Factors

- **Source:** American Community Survey (ACS) – 5-Year Census
- **Description:** Income levels, population density, poverty percentage, community resilience to disasters (CRE) etc
- **Purpose:** Enables analysis of how socio-economic conditions influence or are affected by power outages, and identifies at-risk communities

# Key Variables

## Independent Variables

### Geographical Data

- County with FIPS
- State
- Geo Coordinates (Latitude, Longitude)
- Canopy Cover Percent

### Socio Economic Data

- Population
- Median income
- Population Density
- Poverty percent
- Social vulnerability percent

### Infrastructure

- Overhead transmission line
- Underground transmission line

### Storm Events and Weather

- Event Date
- Event Type: Flash Flood/ Hail/Heat/Heavy Rain
- Event Duration
- Temperature
- Windspeed
- Precipitation

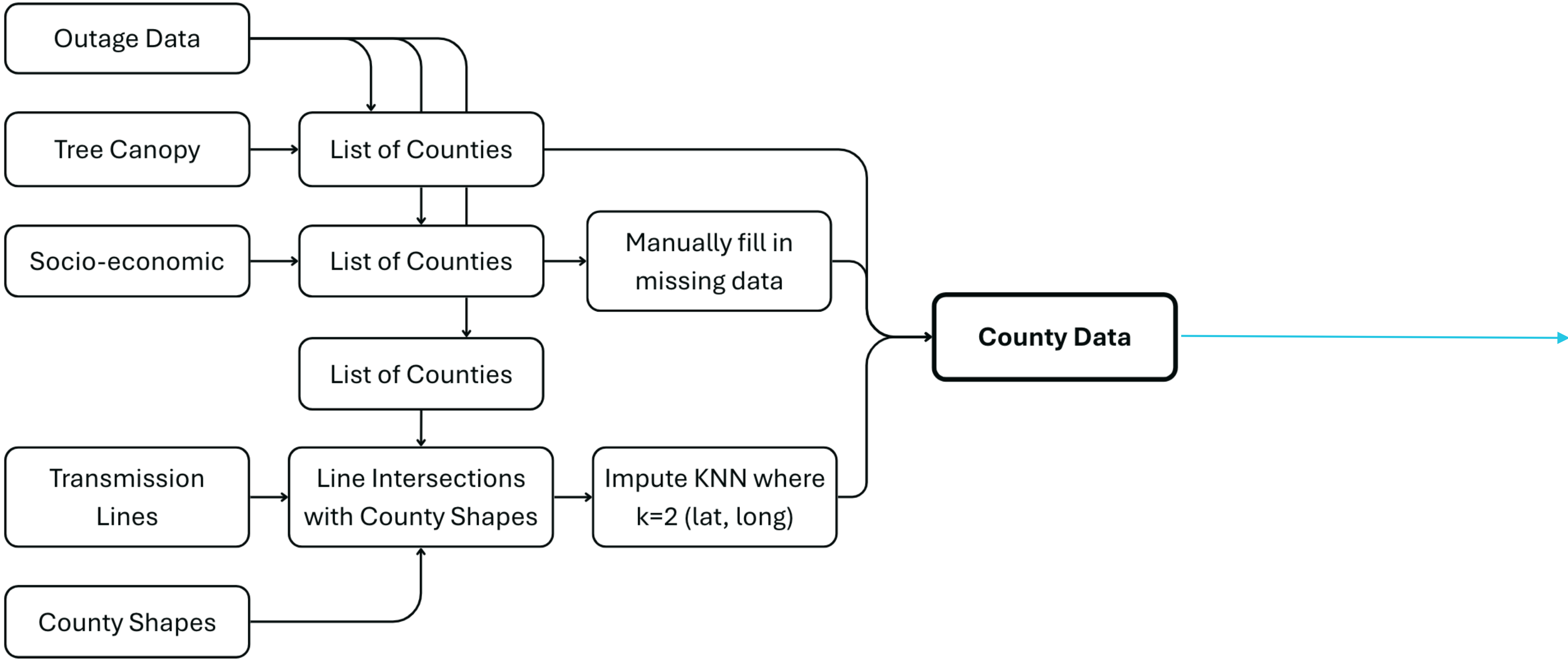
## Dependent Variables

- **Outage Flag**: If outage duration is not zero, Outage Flag = 1
- **Outage Duration** (Minutes)

# Preprocessing

**Step 1:** Generating the county dataset by merging through FIPS code (county identifier)

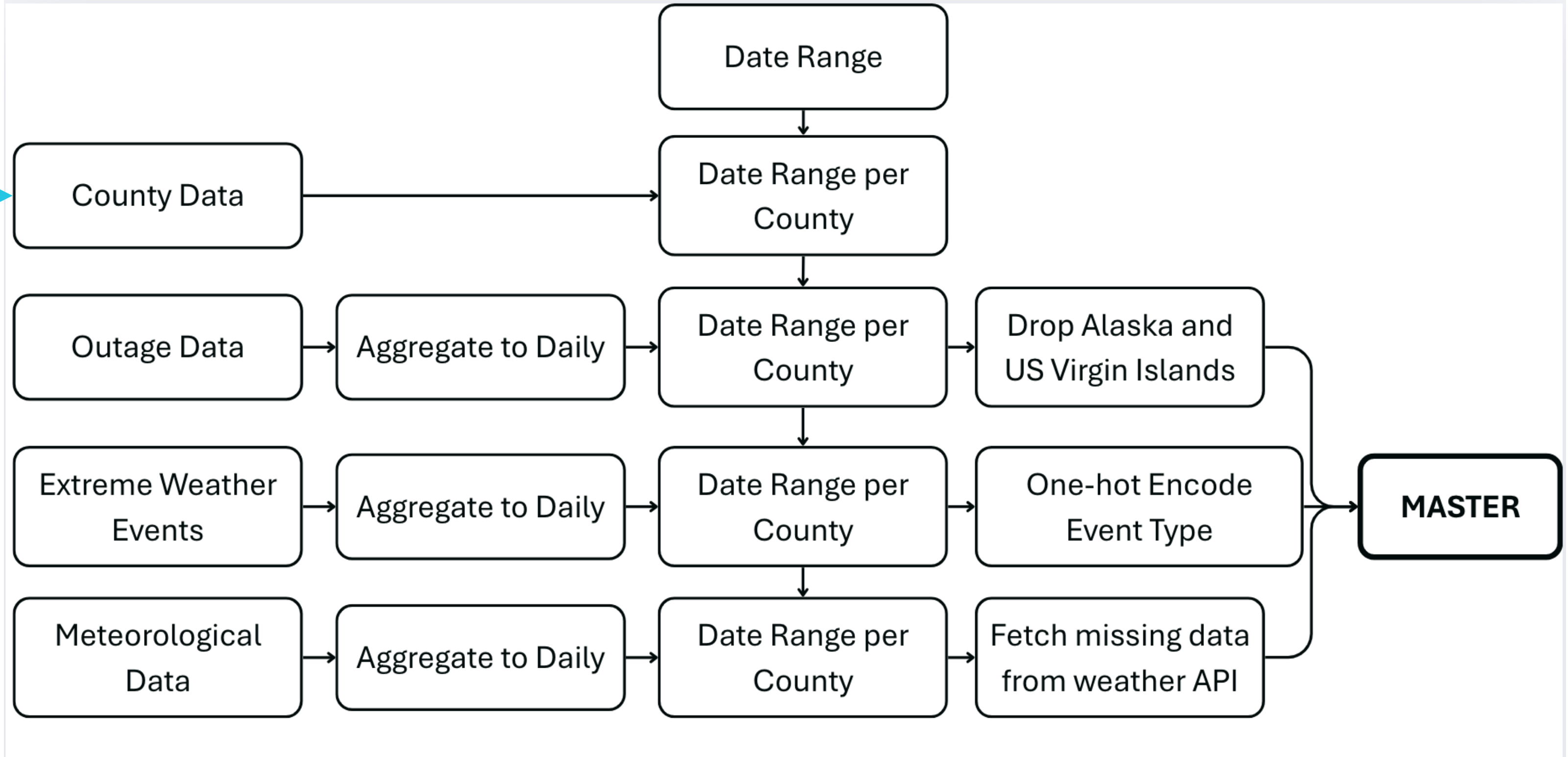
| county_fips | lat | lng | population | canopy_cover_percent | overhead_transmission_line | underground_transmission_line | median_income | poverty_percent | social_ |
|---|---|---|---|---|---|---|---|---|---|
| 45001 | 34.2226 | -82.4593 | 24352 | 0.58092 | 40 | 0 | 51580.0 | 0.15 | |



# Preprocessing (contd.)

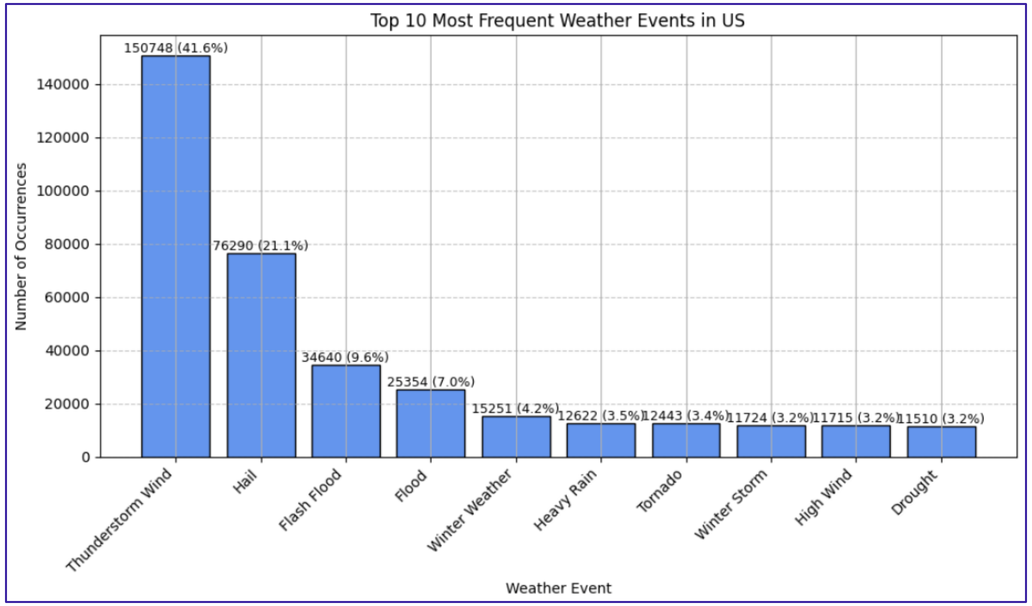
**Step 2:** Generating the table of dates (2014-2023) per county (3041 counties)

| county_fips | date | county | state | lat | lng | population | canopy_cover_percent | overhead_transmission_line | underground_transmission_line | ... | outage_customer_ave | outage_minutes | temp | windspeed | precip |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 45001 | 2014-11-07 | abbeville | south carolina | 34.2226 | -82.4593 | 24352 | 0.58092 | 40 | 0 | ... | 0 | 0 | 11.150833 | 1.549583 | 99.767083 |
| 45001 | 2014-11-08 | abbeville | south carolina | 34.2226 | -82.4593 | 24352 | 0.58092 | 40 | 0 | ... | 4 | 90 | 8.710833 | 1.150833 | 99.744167 |
| 45001 | 2014-11-09 | abbeville | south carolina | 34.2226 | -82.4593 | 24352 | 0.58092 | 40 | 0 | ... | 0 | 0 | 11.725000 | 1.356667 | 99.604167 |



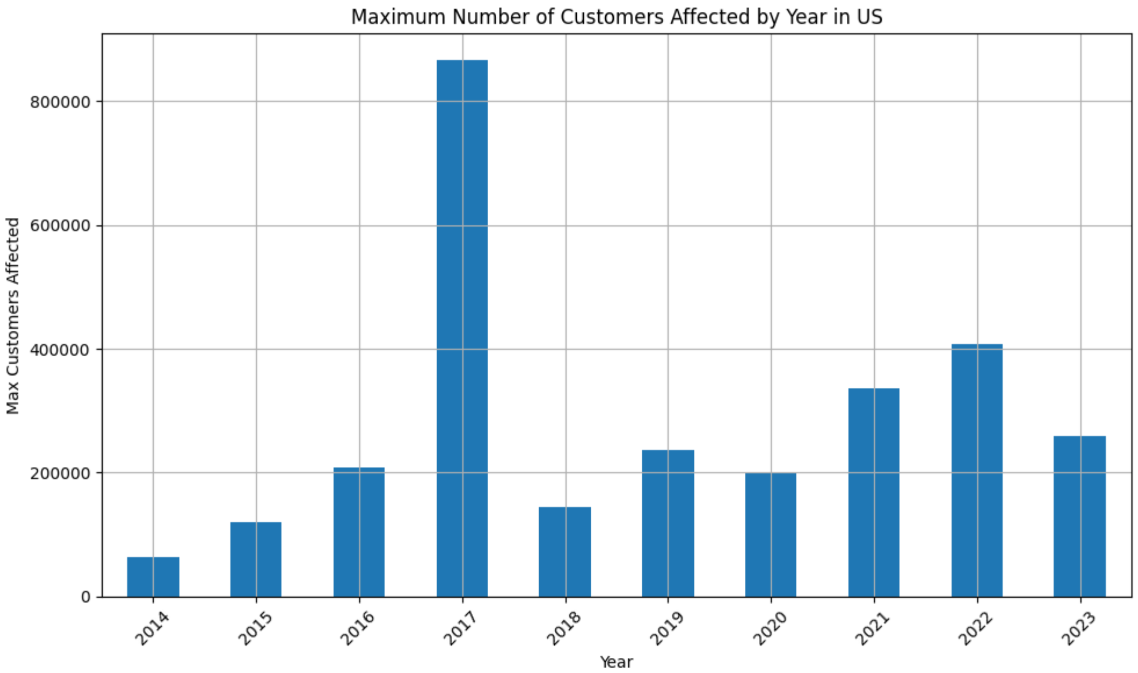
# Exploratory Data Analysis

## Most Frequent Weather Events in US




- Thunderstorm Wind most common weather event by far
- 4 event types accounts for 80% of all extreme weather events
- Steep drop in frequency after the top two categories

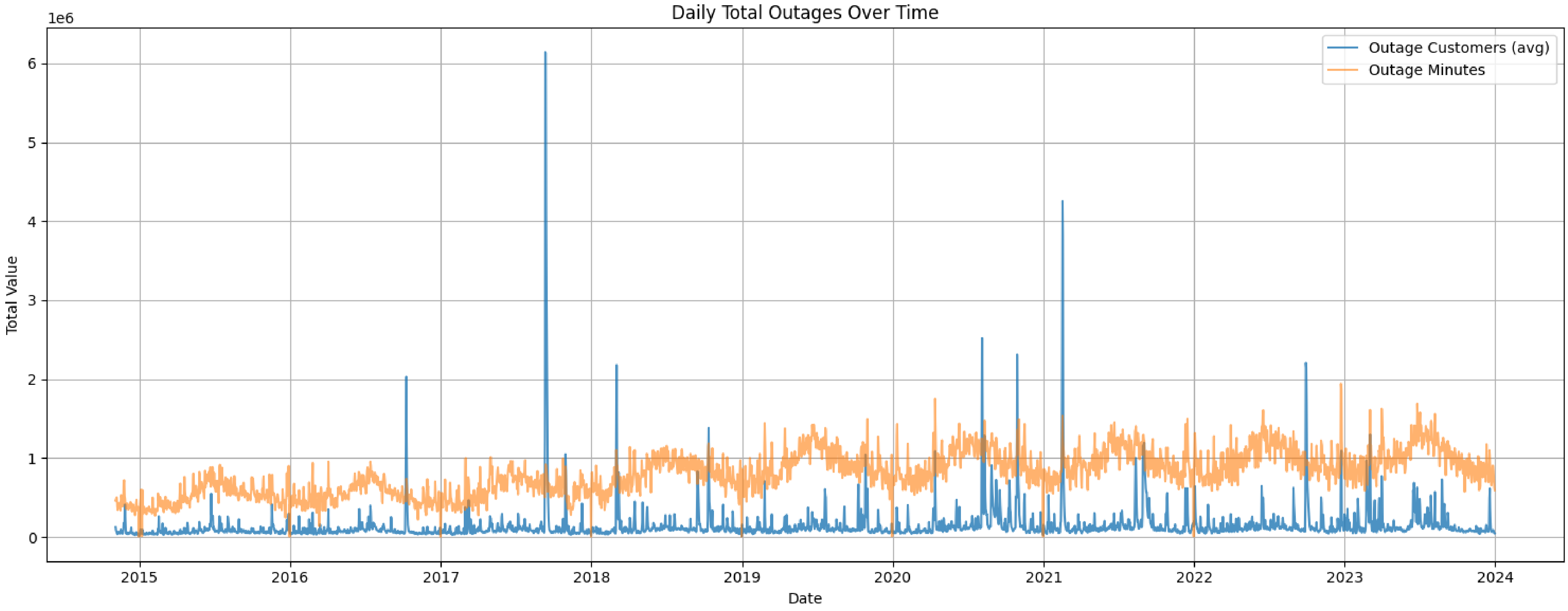
## Number of Customers Affected in US




- 2017 stands out with highest customer impact
- Distribution over the years uneven

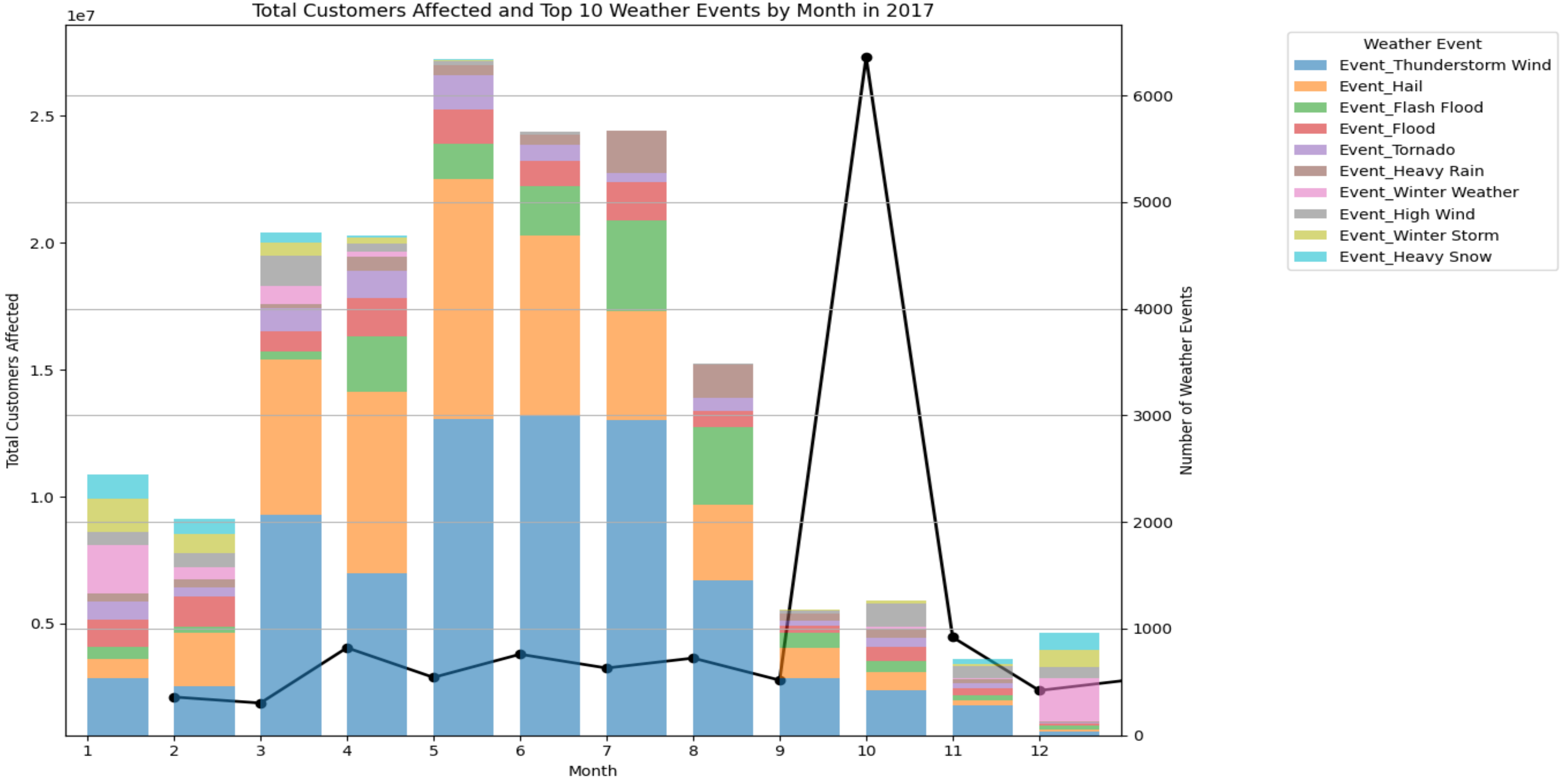
# Exploratory Data Analysis (contd.)

## Daily Total Outages over time



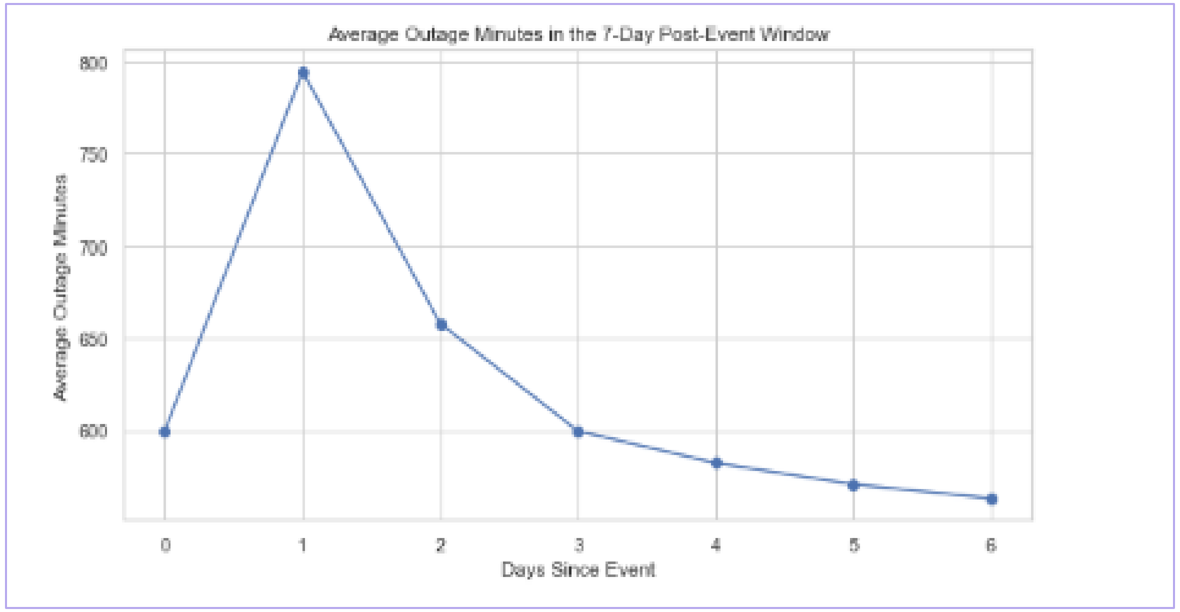
# Exploratory Data Analysis (contd.)

## Top 10 Weather Events by month and Total Customers Affected



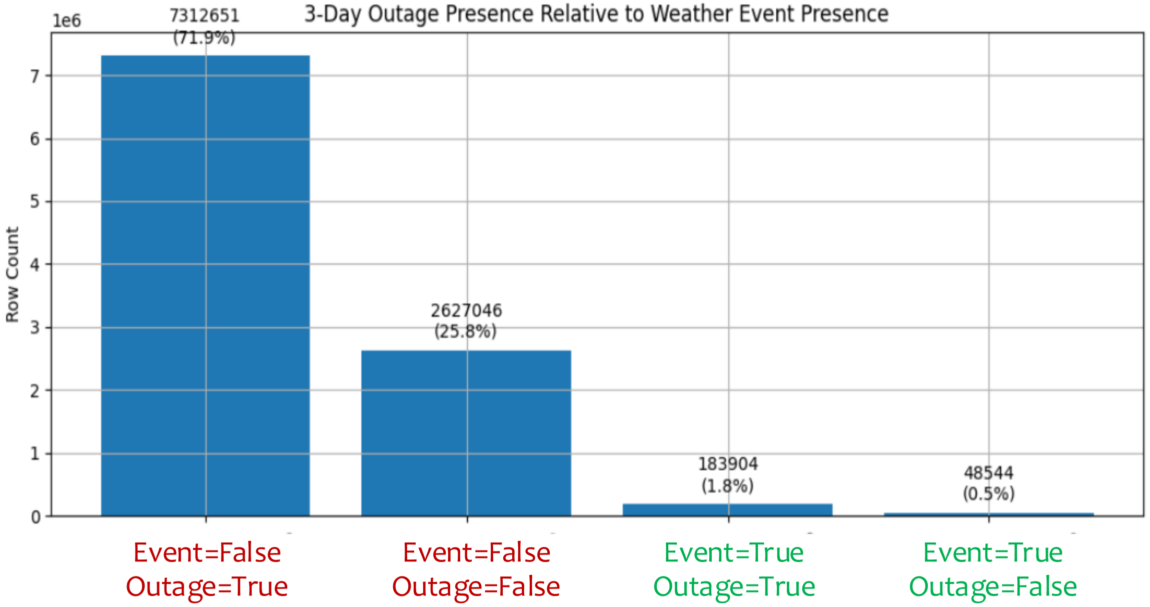
# Exploratory Data Analysis (contd.)

## Choosing ideal post-weather event window based on Impact




- Observed highest outage impact within **3 days** after a weather event.
- So, we introduced lag features for weather/events up to 3 days prior to capture this effect.

## Weather Events vs Outage in next 3 days

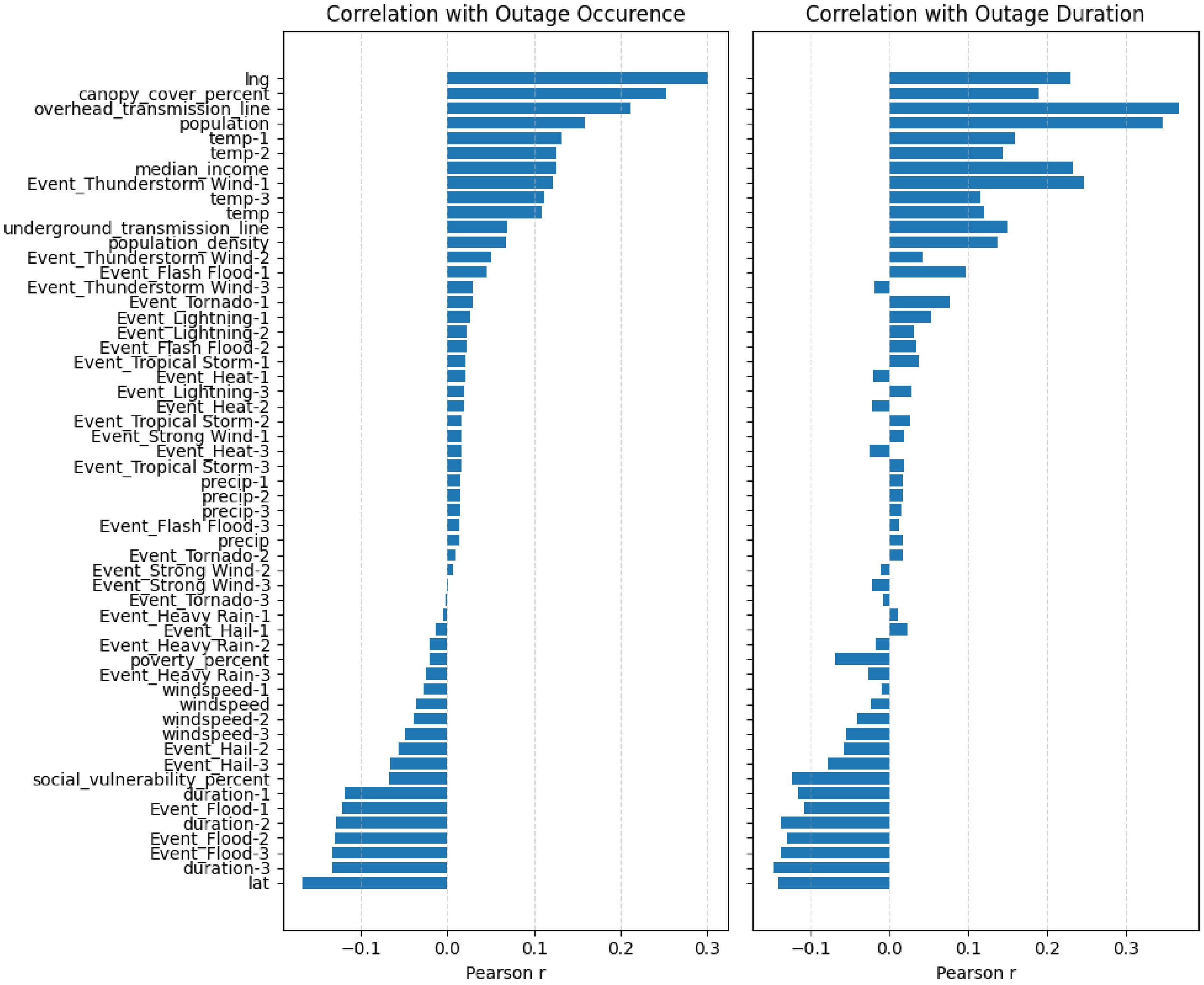



- Only **1.8%** of records had outages which had an extreme weather event in the past 3 days — indicating a weak direct signal.
- To focus on weather-related patterns, we restricted the dataset to 3 days following extreme events in the county

# Exploratory Data Analysis (contd.)

## Correlation Analysis

- Outage Occurence
  - Positive Correlated Features
    - Longitude , Canopy Cover, Overhead Transmission Line, Population, Temp (T-1 day)& (T-2 day)
  - Negative Correlated Features
    - Latitude, Event Duration (T-3), Event Flood (T-3) & (T-2),  Event Duration (T-2)
- Outage Duration
  - Positive Correlated Features
    - Overhead Transmission Line, Population, Event-Thunderstorm(T-1), Median Income, Longitude
  - Negative Correlated Features
    - Event Duration (T-3), Latitude, Event_flood (T-3), Event_Duration (T-2)



# Feature Engineering

### Lagged Features

- Meteorological and extreme weather events: introduced 3-day lagged features to capture recent trends, reflecting short-term meteorological buildup leading to outages.

| Date | Event_Storm | Precipitation |
|---|---|---|
| 2020-01-01 | 120 | 40 |
| 2020-01-02 | 1042 | 50 |
| 2020-01-03 | 0 | 10 |

| Date | Event_Storm-1 | Event_Storm-2 | Event_Storm-3 | Precipitation-1 | Precipitation-2 | Precipitation-3 |
|---|---|---|---|---|---|---|
| 2020-01-04 | 0 | 1042 | 120 | 10 | 50 | 40 |

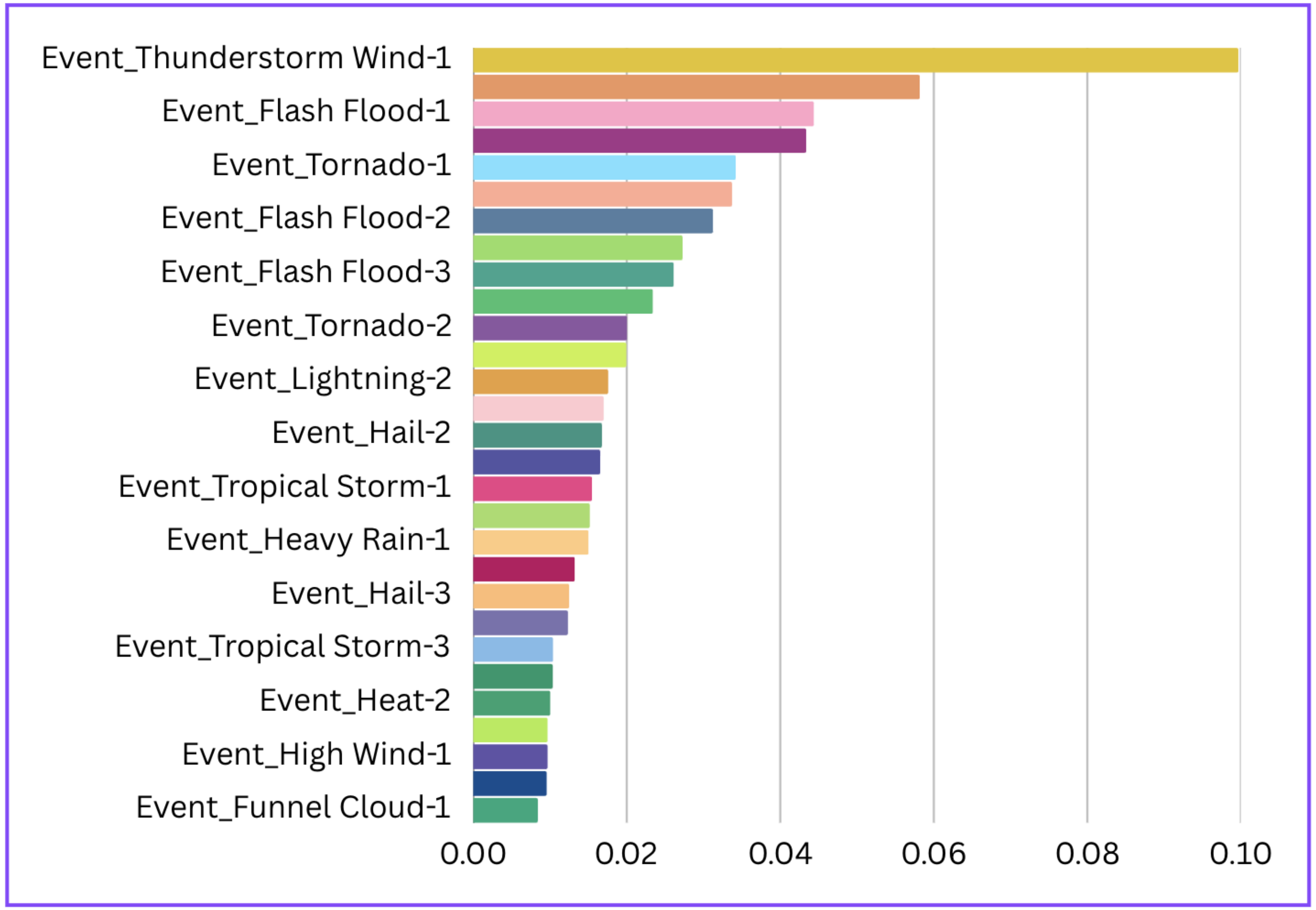
### Filter on Dates with Extreme Weather Event Occurence

- To predict outages driven by extreme weather events

# Feature Engineering (contd.)

## Feature Selection

- Top 30 Weather Events correlated to outage duration (otherwise we would have 165 lagged weather features)



# Feature Engineering (contd.)

**Domain-informed features**

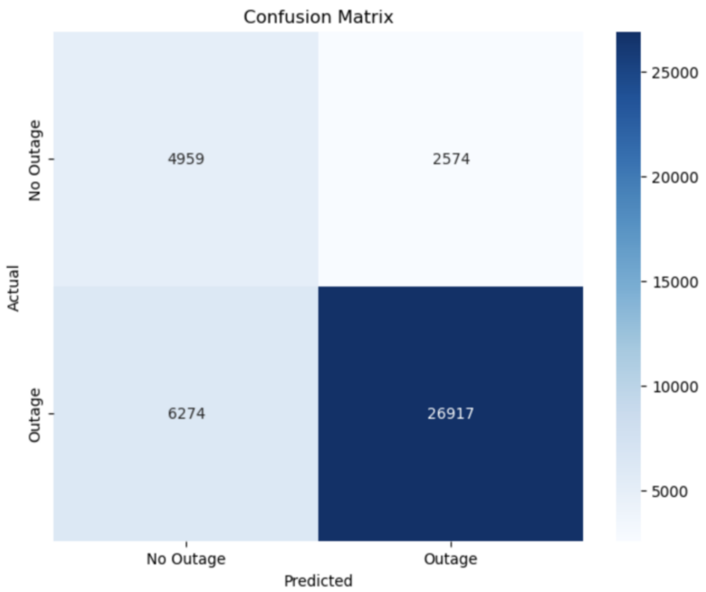
Developed domain-informed features across event, weather, and infrastructure dimensions

- **Event-Based**
  - event_count: Total active weather events.
  - storm_combo, flood_combo: Grouped storm/flood intensities.
- **Weather Summary**
  - precip_total, wind_total: Recent 3-lag weather totals.
  - temp_range: Temperature variability — a proxy for instability.
- **Infra & Social Risk**
  - infra_exposure: Tree cover × overhead lines.
  - social_risk: Poverty × social vulnerability.
  - line_ratio: Underground vs. overhead line mix.
- **Socio Economic**
  - Population Density: County Population / Land Area

# Results: Label 1- Prediction of Outage Occurence (contd.)

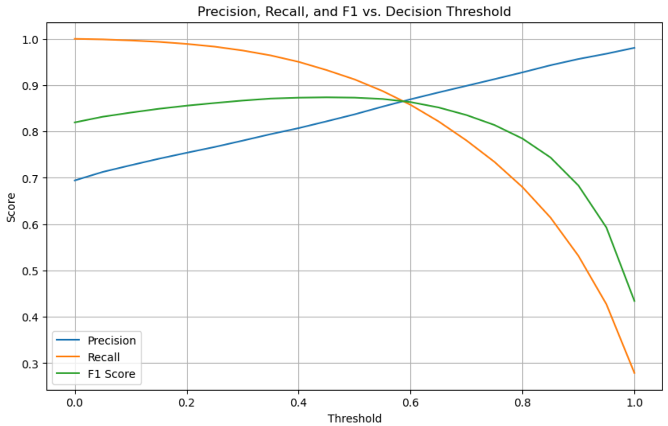
| TRAINING (66%) | VALIDATION (14%) | TEST (20%) |
|---|---|---|

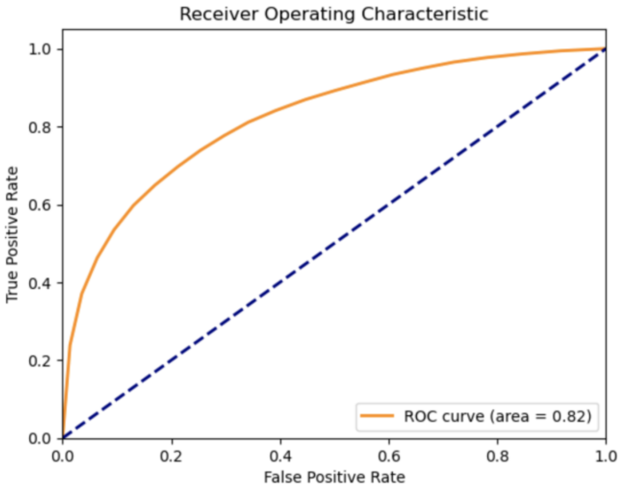
2014-11-07 → 2020-07-13 → 2022-03-10 → 2023-12-31




## Model Validation Performance Results

| Model | Precision | Recall | F1-Score |
|---|---|---|---|
| Logistic Regression | 0.72 | 0.94 | 0.82 |
| XGBoost Classifier | 0.81 | 0.91 | 0.86 |
| Random Forest Classifier | 0.91 | 0.81 | 0.86 |
| Random Forest Classifier (Top 25 Features) | 0.91 | 0.81 | 0.86 |




## Chosen Model : Random Forest Classifier

- RandomForest vs XGBoost: More important to be precise

# Results: Label 1- Prediction of Outage Occurrence (contd.)

| | feature | importance |
|---|---|---|
| 2 | population | 0.096372 |
| 1 | lng | 0.065935 |
| 3 | canopy_cover_percent | 0.054535 |
| 0 | lat | 0.053126 |
| 50 | precip-3 | 0.048021 |
| 44 | precip-1 | 0.046289 |
| 4 | overhead_transmission_line | 0.040634 |
| 42 | temp-1 | 0.045257 |
| 47 | precip-2 | 0.044093 |
| 48 | temp-3 | 0.041353 |
| 45 | temp-2 | 0.041334 |

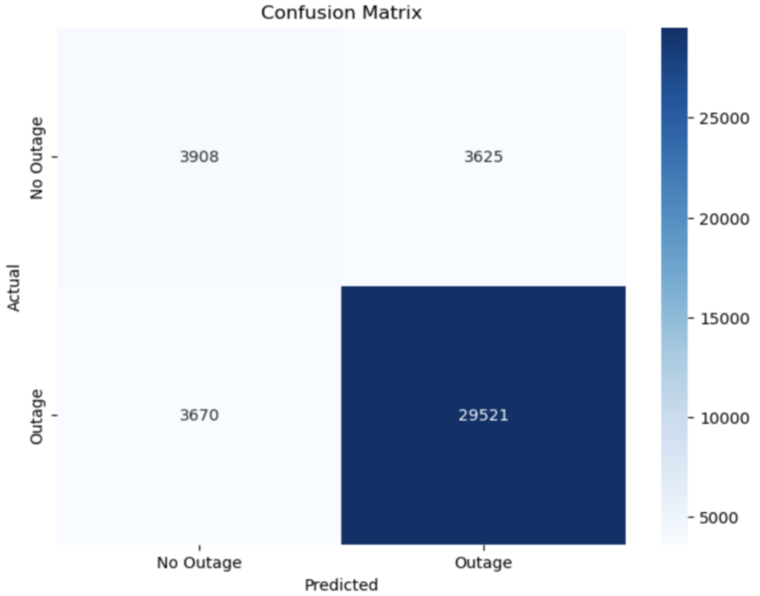
```
# reduce features before tuning
X_train = X_train[important_features]
X_test = X_test[important_features]

n_estimators = [int(x) for x in np.linspace(start = 10, stop = 400, num = 10)]
max_features = ['log2', 'sqrt']
max_depth = [int(x) for x in np.linspace(10, 100, num = 10)]
max_depth.append(None)
min_samples_split = [2, 5, 10]
min_samples_leaf = [3, 4, 5]
bootstrap = [True, False]
```

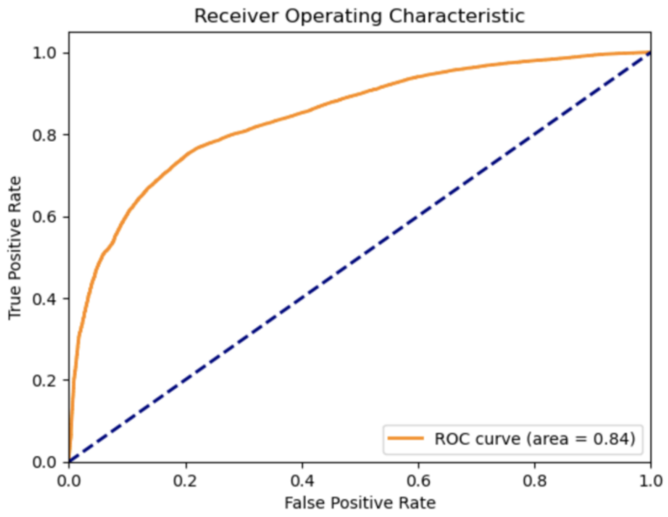
```
[151] rf_random.best_params_

{'n_estimators': 270,
 'min_samples_split': 10,
 'min_samples_leaf': 4,
 'max_depth': 10,
 'bootstrap': True}
```

## Results after Tuning

| Model | Precision | Recall | F1-Score |
|---|---|---|---|
| Validation | 0.89 | 0.89 | 0.89 |
| Test | 0.90 | 0.84 | 0.87 |

- F1-score from 0.86 -> 0.89
- AUC from 0.82 -> 0.84



# Results: Label 2- Prediction of Outage Duration

Comparing shortlisted model performance to identify the most accurate approach for predicting outage duration

## Model Performance Results

| Model | RMSE (min) | $R^2$ |
|---|---|---|
| Linear Regression | 430.546 | 0.272 |
| Random Forest Regressor | 419.069 | 0.391 |
| XGBoost Regressor | 426.290 | 0.370 |

## Best Performing Model : Random Forest Regressor

- Random Forest Regressor achieved the best overall performance:
- Lowest RMSE - most accurate outage duration predictions
- Highest $R^2$ score - explains the most variance in the target variable

## Next Steps: Model Enhancement

- Additional Feature Engineering
- Feature Importance Analysis
- Hyperparameter Tuning

# Results: Label 2- Prediction of Outage Duration (contd.)

## Enhancements

**Additional Feature Engineering**

- Running Average of Outage Duration: Computed average duration of past outages per county to capture local patterns
- Event Duration from storm dataset: Integrated extreme weather event durations as lag features to reflect delayed impact on outages

**Feature Selection via Importance**

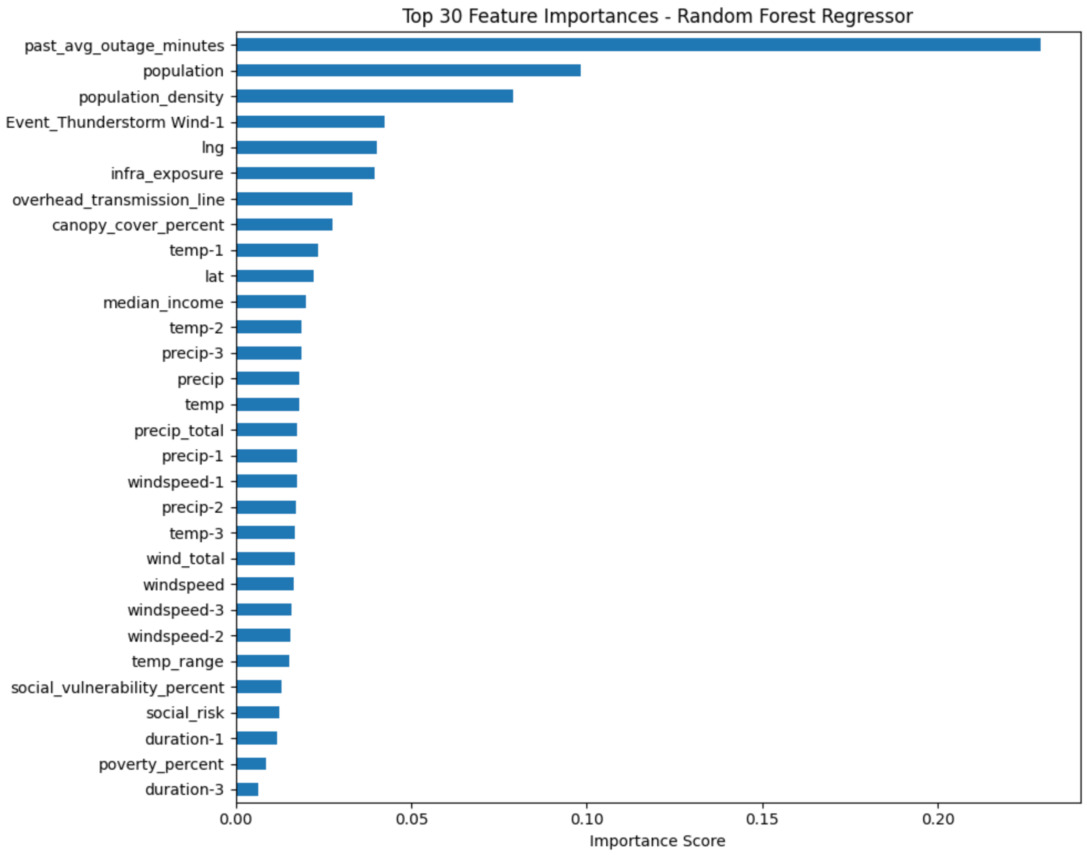
- Extracted feature importance scores and selected top 30 features out of 75, reducing dimensionality and noise.

**Hyperparameter Tuning**

- Performed tuning using RandomizedSearchCV to optimize performance
- Identified best parameter combination:

```
Best Parameters: {'bootstrap': True, 'max_depth': None,
'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_sampl
es_split': 4, 'n_estimators': 262}
```

## Feature Importance Plot



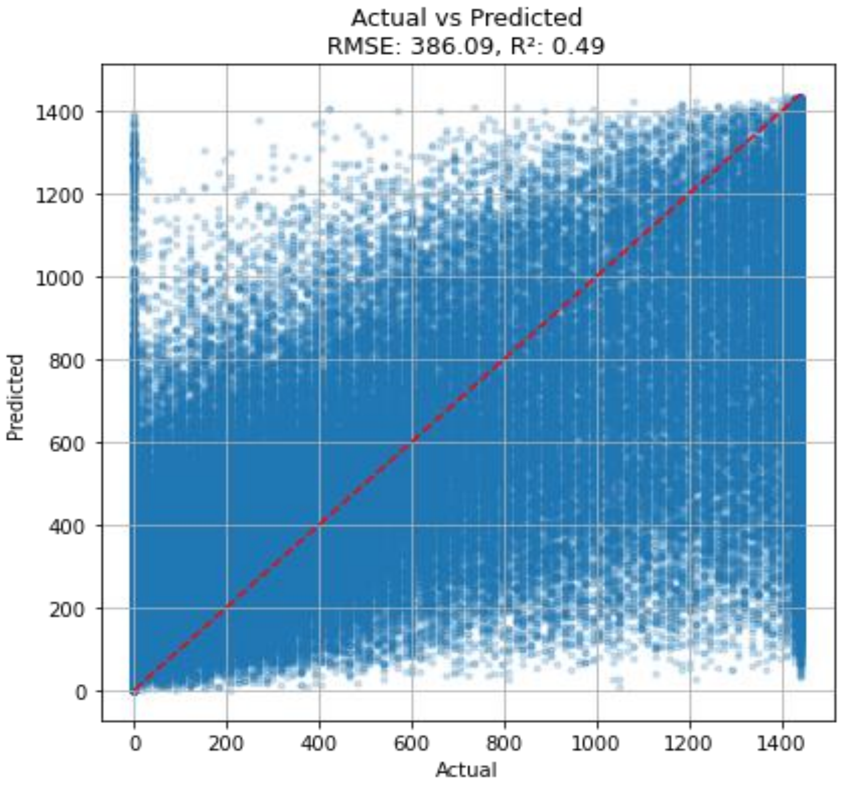
# Results: Label 2- Prediction of Outage Duration (contd.)

Improvements driven by targeted feature selection and model optimization techniques

### Enhanced Model Performance

| Random Forest Regressor Model | RMSE (min) | $R^2$ |
|---|---|---|
| Feature Engineering & Selection | 386.30 | 0.484 |
| After Hyper Tuning | 386.09 | 0.485 |

- Feature Engineering and Selection
  - RMSE reduced from 419 → 386 (~8% reduction)
  - $R^2$ improved from 0.391 → 0.484, showing greater explanatory power
  - No of independent features reduced from 65 to 30
- Hyperparameter Tuning
  - RMSE further reduced to 386.09
  - $R^2$ improved from 0.484 → 0.485



# Challenges

### Challenges During Execution

- **Data Integration Complexity:** Aggregated ~10M records, creating preprocessing challenges. Filtering relevant data for EDA and ML was difficult
- **Event vs Outage** : Few outages were clearly linked to extreme weather events, making signal detection hard
- **Additional Labels:** Despite efforts to predict an additional label- 'affected customers', but feature importance analysis showed weak dataset signals.
- **Lag Features:** Struggled to define how to effectively create lagged features to represent the build-up to outages, given multiple types of weather events involved.

### Lessons Learnt

- Model performance depends heavily on data quality. Without clean, reliable data, models are ineffective.
- Majority of time spent on EDA and preprocessing.
- Understanding data context is key to insights.
- Limited compute resources can slow project development significantly

### Areas for Improvements

- **Refine Prediction Granularity for Real-Time use:** Shift modeling predictions to minutes or hours before events instead of daily estimates for timely interventions
- **Better definition of outage:** Defining weather-induced outages is challenging due to many low-customer-impact events : requires domain expertise to validate

# THANK YOU