Weather Severity Prediction Report

1. Introduction

Stakeholder

Our primary stakeholder is the Local Government Emergency Management Team, responsible for monitoring weather conditions and managing emergency responses. Their mission is to leverage data-driven insights to improve the readiness and effectiveness of emergency services during extreme weather events.

Problem Statement

Extreme weather events—such as heavy rain, snowstorms, and thunderstorms—pose significant risks to public safety and infrastructure. The challenge is to develop a predictive model that forecasts the severity of these weather events (rated on a scale of 1 to 4) using historical data. An accurate prediction model can help the team plan resource allocation, issue timely warnings, and ultimately reduce the adverse impact of severe weather on communities.

---

2. Dataset Overview

Dataset Source

The dataset comprises historical weather records collected from various airports and weather stations. It includes:

- Precipitation (in inches)
- Event Start and End Times
- Geographical Data (Latitude, Longitude)
- Event Type
- Event Severity (the target variable)

Dataset URL : https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

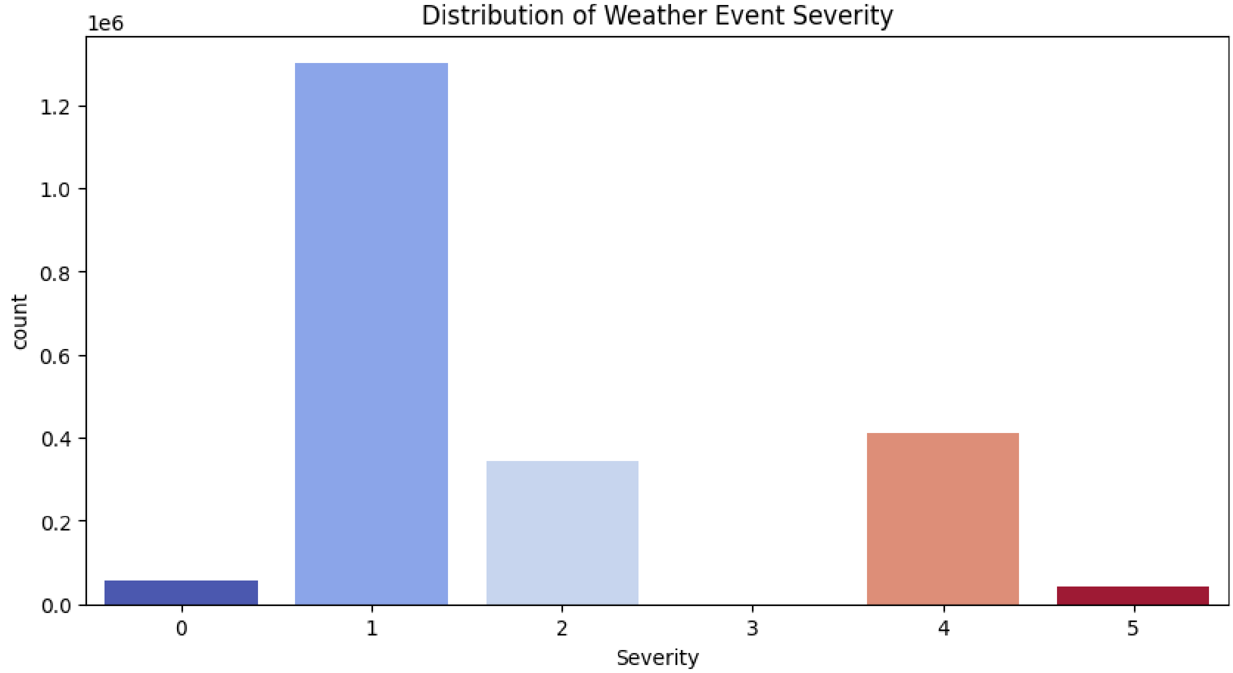
Key Attributes and Visualizations

- Precipitation: Measures the intensity of rainfall or snowfall.
- Event Duration: Derived by calculating the difference between event start and end times.

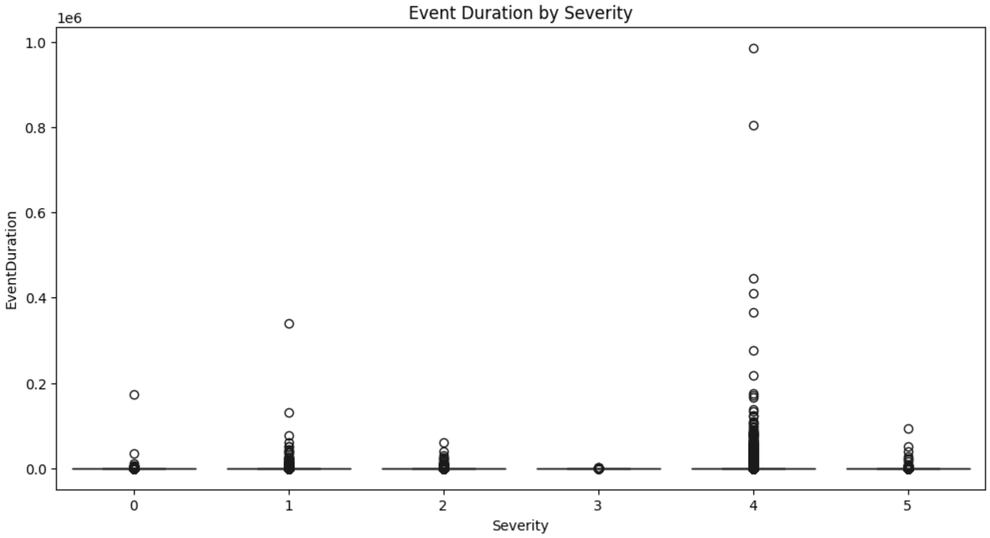
- Temporal Features: Month, hour of day, and day of the week, capturing seasonal and daily patterns.
- Categorical Features: Event type and severity (encoded for modeling).

Visualizations created during exploratory analysis include:

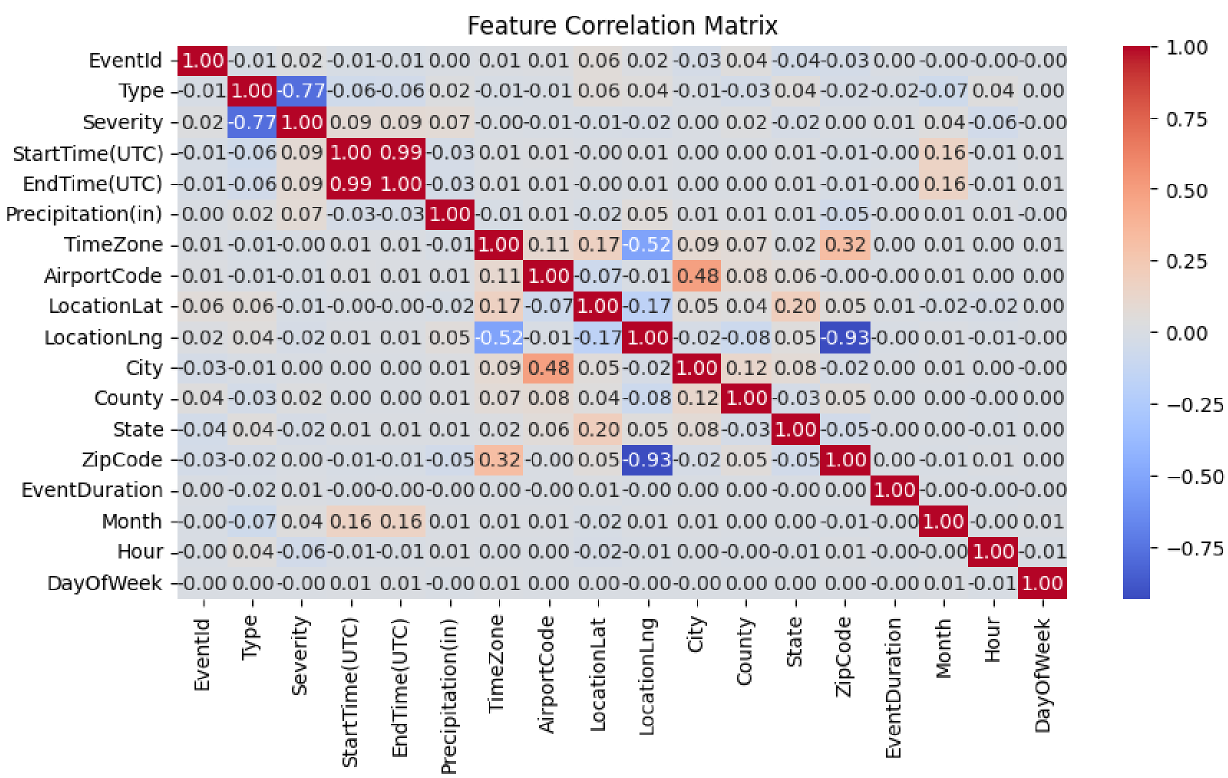
- **Severity Distribution Plot:** A count plot illustrating the frequency of each severity level.



- **Boxplot of Event Duration by Severity:** Shows the distribution and variance of event durations.



- **Correlation Heatmap:** Highlights relationships among numerical features, supporting feature selection.



3. Feature Engineering

Selected Features

The core features used in our model include:

- **Precipitation (in inches)**
- **Event Duration (in minutes)**
- **Month:** Extracted from the event timestamp.
- **Hour of Day:** Reflects diurnal patterns.
- **Day of the Week:** Captures weekly cyclic trends.
- **Event Type:** Categorical representation of weather events.

Engineered Features

To enhance model performance and capture complex interactions, we engineered additional features:

1. **Weather Severity Index:**
   A composite index combining precipitation and event duration. This normalization helps quantify the overall intensity of an event more robustly than individual features.
2. **Time-of-Day Impact:**
   Hours were binned into categories (morning, afternoon, evening, and night) to capture varying impacts of weather events at different times.
3. **Seasonality Indicator:**
   A binary feature indicating whether the event occurred during peak seasonal months (e.g., winter for snowstorms), further refining the model's seasonal sensitivity.

Rationale

- **Direct Indicators:** Precipitation and duration provide immediate signals about the intensity of a weather event.
- **Temporal Patterns:** Month, hour, and day capture recurring patterns that can influence severity.
- **Engineered Features:** The Weather Severity Index and Time-of-Day Impact offer deeper insights into event characteristics that might be missed when using raw features alone.

4. Model Selection and Training

Models Evaluated

Two models were compared to determine the best solution:

1. **Random Forest Classifier**
   - **Pros:**
     - Handles complex, non-linear interactions well.
     - Robust against overfitting with its ensemble approach.
     - Provides feature importance, aiding interpretability.
   - **Cons:**
     - Computationally demanding with large parameter grids.
   - **Hyperparameter Tuning:**
     - Parameters tuned included n_estimators, max_depth, min_samples_split, and min_samples_leaf.
     - Over 24 parameter combinations were tested using cross-validation.
2. **XGBoost Classifier**
   - **Pros:**
     - Excellent performance on structured/tabular data.
     - Efficient training and effective handling of imbalanced datasets.
   - **Cons:**
     - More sensitive to hyperparameter selection.
     - More complex to interpret compared to Random Forest.
   - **Hyperparameter Tuning:**
     - Parameters tuned included n_estimators, learning_rate, max_depth, and subsample.
     - A similar grid search process was applied.

Tuning Results and Final Selection

The performance of the models was evaluated as follows:

- **Random Forest Accuracy:** 0.9142
- **XGBoost Accuracy:** 0.9140

Random Forest was marginally better, and thus selected as the final model.

5. Model Evaluation

Evaluation Metrics

- **Accuracy:**
  The primary metric indicating overall correctness of predictions.
- **Precision & Recall:**
  Assessed to ensure that severe weather events are correctly identified with minimal false positives.
- **Confusion Matrix:**
  Visual representations confirmed that most predictions fall along the diagonal, indicating strong model performance.

Visual Insights

- **Confusion Matrix Plot:**
  Reveals that misclassifications are minimal, reinforcing model reliability.
- **Severity Distribution and Boxplots:**
  Provide qualitative insights that complement quantitative metrics, aiding in feature refinement and model interpretation.

6. Future Work & Recommendations

Future Enhancements

- **Additional Data:**
  Integrate more meteorological variables (e.g., wind speed, humidity) to further enhance prediction accuracy.

- **Advanced Models:**
  Explore deep learning models such as LSTM networks for capturing temporal dependencies more effectively.
- **Model Interpretability:**
  Apply SHAP (SHapley Additive exPlanations) to understand feature contributions and improve transparency.
- **Real-Time Updates:**
  Implement a pipeline for continuous model retraining with new data to maintain performance over time.

Recommendations for Deployment

Based on the rigorous evaluation:

- **Final Model:** RandomForestClassifier
- **Performance:** Achieved an accuracy of 91.42%, with precision and recall values that confirm its reliability.
- **Deployment:**
  The model is suitable for integration into a web-based interface (using tools like Flask, Gradio, or Streamlit) to provide real-time predictions.
- **Stakeholder Benefit:**
  The model can significantly enhance emergency preparedness by providing timely and accurate severity predictions, ultimately aiding in resource allocation and public safety decisions.

7. Conclusion

This project demonstrates a successful application of machine learning techniques to predict weather severity. By combining robust feature engineering, careful model selection, and thorough evaluation, we achieved a highly accurate model. The selected Random Forest model, with an accuracy of 91.42%, is both reliable and interpretable, making it a valuable tool for the Local Government Emergency Management Team.

The insights drawn from extensive visualizations further validate the model's capability and provide a roadmap for future improvements. Continuous updates and enhancements will ensure that the model remains a critical asset in mitigating the impacts of severe weather events.

Also, I have tried to deploy the model using Gradio package in google colab. The mode works well for clear and cloudy data.