Storm Damage Prediction Using NOAA Storm Events Data

Machine Learning–Based Estimation and Feature Attribution of Property and Crop Losses caused by Desasters

Nathania Gabriela
Duc-Anh Nguyen

October 2025

1 Introduction

Extreme weather events such as floods and hurricanes lead to extensive economic losses and damage towards property and crops. This project develops a machine learning (ML) pipeline to predict property and crop damages from storm events based on structured data and narratives. We also wish to explain which features influence these damages.

2 NOAA Dataset

The **NOAA Storm Events Database** [NCEI(2025)] provides detailed records of significant weather and climate events across the United States, including event type, location, duration, magnitude, and reported impacts Each entry corresponds to a storm event documented by the National Weather Service, with accompanying information on property and crop damages, fatalities, and narrative descriptions of impacts. The dataset spans 1950 to the present and covers all U.S. states and territories at county-level resolution. It serves as a key source for modeling storm losses, assessing climatological trends, and developing probabilistic damage prediction models.

3 Regression Models

This section outlines the modelling approaches utilised. It estimates the damaged properties and crops based on the given data.

3.1 Ridge regression

Ridge regression is a linear modeling technique that incorporates L2 regularisation to prevent overfitting and improve generalisation [Hoerl and Kennard(1970)]. The model mimimises

$$L(\beta) = ||y - X\beta||^2 + \alpha||\beta||^2,$$

where y represents the true value, X the feature matrix, β the regression coefficients, and α the regularisation strength.

3.2 eXtreme Gradient Boosting (XGBoost Regression)

XGBoost implements a gradient-boosted decision trees (GBDT) [Chen and Guestrin(2016)]. A GBDT is a decision tree ensemble learning algorithm. GBDT trains on a weak decision trees. It would then additively built the next trees based off of the base learner's error residual. The residuals are aggregated to evaluate the model, where the final prediction is the weighted sum of all tree predictions. It minimises the following regularised objective at iteration t

$$L^{(t)} = \Sigma_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t),$$

where l is a differentiable loss function and $\Omega(f_t)$ is a regularisation term.

3.2.1 XGBoost with Quantile Loss

When XGBoost uses quantile loss as its objective, it estimates the conditional median or quantiles of the target distribution instead of the conditional mean. The quantile loss is defined as

$$L_\tau(y,\hat{y}) = \begin{cases} \tau(y-\hat{y}), & \text{if } y > \hat{y} \\ (1-\tau)(y-\hat{y}), & \text{otherwise} \end{cases} \tag{1}$$

where $\tau \in (0,1)$ is the quantile of interest.

3.2.2 XGboost with Tweedie Distribution

Tweedie distribution is a general family of exponential dispersion models (EDMs), which includes Gaussian, Poisson, Gamma, Compound Poisson-Gamma, and more. EDMs constitutes of

$$f(y|\theta,\phi) = h(\phi,y)exp(\frac{\theta y - b(\theta)}{\phi}),$$

where θ is the canonical parameter, $\phi > 0$ is the dispersion parameter, and $b(\theta)$ determines the mean-variance relationship. The variance is defined as $Var(Y) = \phi V(\mu)$. The Tweedie distribution introduces power parameter $p \in \mathbb{R}$, which variance is now $V[Y] = \phi V(\mu) = \phi\mu^p$ [Tweedie(1984)]. The Tweedie distribution corresponds to Gaussian distribution when $p = 0$, Poisson distribution when $p = 1$, Gamma distribution when $p = 2$, and Compound Poisson-Gamma when $1 < p < 2$ [Dunn and Smyth(2005)]. The probability density function of x is

$$f(x \mid \mu,\phi,p) = a(x,\phi,p)\cdot\exp\left\{\frac{1}{\phi}\left(x\cdot\frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right\}.$$

The Tweedie loss is derived from the negative log-likelihood, which is

$$L(\mu) = -x\cdot\frac{\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}.$$

4 Experiment

4.1 Preprocessing: EDA and Feature Engineering

We evaluated our models using the NOAA Storm Events Database, focusing on records from the year 2013 [NCEI(2025)]. The raw datasets contained numeric, categorical, and textual data. Firstly, we remove duplicates and irrelevant variables. We aggregated some of the features, such as the beginning and ending of the event, into a duration feature. Then, we compute the mean latitude and longitude along with the haversine distance. We also compute the duration of the event in hours.

We then handled the missing values according to their data types. For numeric features, we impute missing values with the median, and then we use `StandardScaler` function to standardise the data. For categorical features, we impute the missing values with the mode and one-hot encode the features. For the text features, we impute the missing values

with an empty string and embed the strings using a pre-trained sentence-transformer model `paraphrase-MiniLM-L6-v2`. We also transformed the target variable with `log1p`.
Then, we split the dataset into training, validation, and test sets with a holdout split. We use 0.7 of the dataset as training and 0.15 for both the validation and test set. We fill the target variables with 0 when values are missing. We impute the value of 0, because we deemed it unlikely for damages not to be recorded.

4.2 Training and Evaluation

We fit multiple models. Firstly, we fit ridge regression and XGBoost as a baseline. Ridge regression was used as a linear baseline model, while XGBoost was used as a nonlinear ensemble method, which cabn model complex interactions and nonlinear relationships. We evaluate them based on the metrics of MAE, RMSE, and R-squared. We fit the XGBoost with quantile loss. This is done to capture the behaviours across different quantiles. We measure them through quantiles 0.1, 0.5, and 0.9 with RMSE and pinball loss. Then, we tune and fit XGBoost with the Tweedie distribution. Tweedie distribution suited for target variables that is non-negative and continuous. It is also suitable for highly skewed and zero-inflated data, which is suitable for our data set. We tuned it to $1 < p < 2$ as it can capture both zero and non-zero events. Finally, we calculate the SHAP values of the features.

5 Results

5.1 Feature Importance and Model Performance Summary

5.1.1 Top 20 Features for Property vs. Crop Damage Predictions

The property damage model's top feature is text-derived (`EventNarr_emb319`), indicating that certain keywords in the event narrative strongly influence property loss predictions. As it is indicated, being the feature with the highest mean SHAP value of property damage in figure 1.
In contrast, the crop damage model's top predictors are contextual "neighbour event" features, e.g., number of nearby or recent events, suggesting that clustering of events in space or time is a key indicator for crop damage. Notably, event duration and event span appear among the important features for both targets, implying that longer-lasting or larger-scale events tend to increase both property and crop damages.
However, there are clear differences: for example, injuries are a prominent feature for property damage, whereas geographic location and specific event types like drought are critical for crop damage predictions.

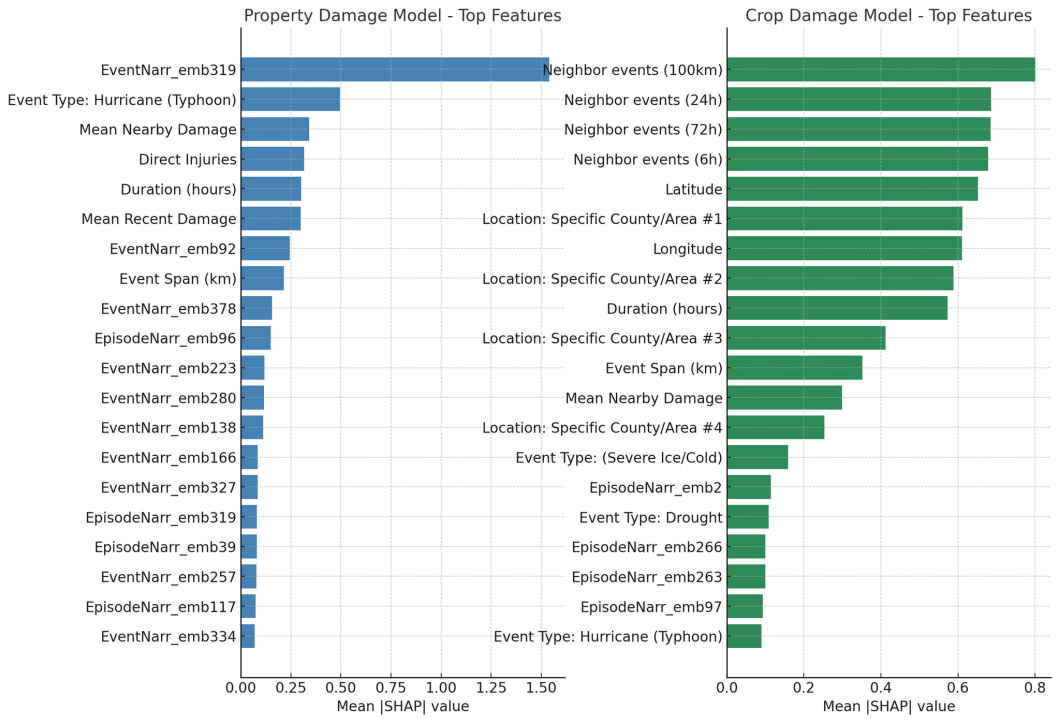

Figure 1: Top 20 SHAP features for the **XGBoost (Tweedie)** property and crop damage models. The left panel displays the top 20 features influencing property damage predictions, while the right panel shows those for crop damage. In both plots, the x-axis represents the mean absolute SHAP value, and the y-axis lists the top-ranked features contributing to model predictions.

5.1.2 Feature Importance for Property Damage Model

Table 2 lists the top 20 features for the property damage model, along with their mean absolute SHAP value, a measure of global importance. Higher SHAP magnitudes mean the feature has a larger influence on the model's predictions on average. We can see that many of the top features are related to the event narrative text (embedded text features) and event severity indicators.

Several of the highest-ranking predictors are **embedding dimensions derived from the event narratives** (e.g., `EventNarr_emb319`, `EventNarr_emb92`). This indicates that the textual descriptions accompanying storm reports contain information that is predictive of damage severity. Because these embeddings are latent numerical representations rather than explicit keywords, their individual semantic meaning cannot be directly interpreted. Nevertheless, their consistent prominence among the top SHAP features suggests that the model captures useful contextual cues from the narratives—such as tone, structure, or co-occurring terminology—that are correlated with more severe events. Further interpretability work (for instance, token-level SHAP analysis or exemplar text inspection) would be required

to determine which linguistic patterns are most influential.
The second-most important feature is the event type being **Hurricane (Typhoon)**, which intuitively aligns with expectations. Hurricanes tend to cause extreme property damage, and the model gives a big boost whenever an event is identified as a hurricane or a cyclone.
We also see **Mean Nearby Damage** and **Mean Recent Damage**. These are aggregate features of past events (e.g., average damages in nearby locations or recent times). It indicates that the model considers historical context.
Direct injuries appear at rank 4, which can serve as a proxy for event severity.
Classic meteorological or temporal features like **event duration** and **event span (area)** are also in the top 10, meaning longer-lasting and wider-ranging events generally lead to greater property damage.

5.1.3 Feature Importance for Crops Damage Model

In Table 3 are the top 20 features for the crop damage model, with their mean |SHAP| importance. This model's important features are noticeably different, emphasizing environmental context and specific event types that affect agriculture.
The top four features are **neighbour event counts**. It indicates how many other events are occurring in proximity, spatially within 100 km, and temporally within the last 6 to 72 hours. All four have high importance, showing that crop damage is often associated with widespread or clustered events. The model also relies on location features and specific county indicators, confirming that certain regions are far more susceptible to high crop losses.
We also see **Drought** identified as an important event type, while hurricanes appear less important for crops than for property.

5.2 Comparing Important Features Between Property and Crop Models

There is some overlap in top predictors between the two targets, but also clear differences reflecting the nature of what drives each type of damage.

- **Event Severity vs. Event Frequency:** The property damage model leans on features that indicate the severity or magnitude of a single event. In contrast, the crop model emphasizes the frequency and clustering of events.
- **Text vs. Location:** Property predictions rely heavily on narrative text features, whereas the crop model depends more on geographic identifiers.
- **Common factors:** Both include event duration, span, and contextual damage history, but with different emphasis.

In summary, property damage is driven by intense, localized events, while crop damage arises from regional or prolonged conditions such as drought or frost.

5.3 Interpreting Quantile Regression Predictions (P10, P50, P90)

Figure 2 and Figure 3 evaluate the calibration quality of the quantile regression models. These diagnostics assess whether the predicted quantiles (*p10*, *p50*, *p90*) correspond to the correct empirical probabilities and whether the uncertainty intervals are appropriately wide.

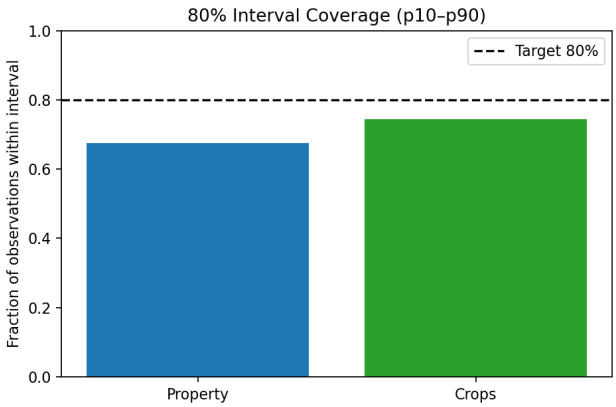

Figure 2: 80% interval coverage for property and crop damage models. The dashed line indicates the ideal 0.8 target.

80% Interval Coverage: The coverage chart (Figure 2) shows the fraction of observed damages that fall within the predicted 10th–90th percentile range—the model's nominal 80% interval. The dashed line marks the ideal target of 0.8. In our results, approximately 68% of property damage outcomes and 75% of crop damage outcomes fell inside this range. This means both models slightly **under-cover**: their predicted intervals are somewhat too narrow, implying that the models are mildly *overconfident* about the precision of their forecasts. In practice, this indicates that extreme damages occur slightly more often than the models expect.

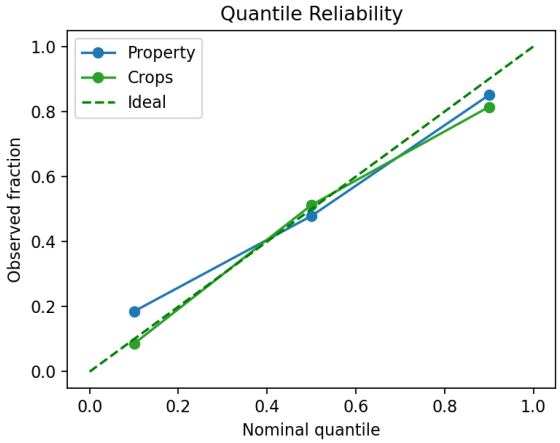

Figure 3: Quantile reliability (calibration) plot for property and crop models. Points close to the diagonal indicate well-calibrated quantile predictions.

Quantile Reliability: The reliability plot (Figure 3) compares each predicted quantile level to the actual proportion of observations below that threshold. A perfectly calibrated model would follow the 1:1 diagonal (dashed green line). Both property and crop models follow the diagonal closely, with minor deviations. The *p10* point of the property is slightly above the ideal line, while the *p90* point of both the property and crops is slightly below. This pattern indicates that the models are **well-calibrated overall**, with a small tendency to underestimate uncertainty at the lower and upper tails.

Interpretation: Together, these results confirm that the quantile models produce **credible probabilistic forecasts**. Their predicted percentiles correspond closely to observed frequencies, meaning statements such as "there is an 80% chance that damages will fall within this range" are generally reliable. The slight under-coverage suggests the uncertainty intervals could be marginally widened to achieve perfect calibration. Overall, the quantile regression framework provides interpretable and actionable *risk ranges* rather than single deterministic estimates, offering valuable insight for uncertainty-aware decision-making.

5.4 Interpreting Model Performance Metrics

Table 1: Model performance metrics for property and crop damage prediction.

Model	R^2	**MAE ($)**	**RMSE ($)**
Property Damage	0.058	109,925	2,646,976
Crop Damage	0.191	46,407	2,358,343

The low R^2 values suggest limited explanatory power, indicating that much of the variance in damage outcomes remains unexplained. High RMSE values are driven by rare, large-magnitude events that the model fails to capture precisely. In contrast, the relatively moderate

MAE to the RMSE implies that, for the majority of cases, the model produces stable predictions close to the observed values, with errors dominated by infrequent extreme losses.

5.5 Discussion

- **Drivers of Damage:** Property damage stems from severe, isolated events; crop losses from widespread climatic patterns.
- **Prediction Uncertainty:** Quantile intervals communicate risk bands, not false precision.
- **Model Accuracy:** Weak—use predictions as guidance, not absolute forecasts.

6 Conclusion

The models provide interpretable insights but limited predictive power. Future refinements should target improving R^2, reducing large-error outliers, and enhancing quantile calibration.

Limitation and Future Work
Some feature variables, such as magnitude and magnitude type, are dropped as more than half of the values are missing.
A method to fix this problem is to impute values through regression.
While embedding features demonstrated strong predictive importance, time limitations prevented a detailed examination of their semantic content. Therefore, the specific meanings represented by individual embedding dimensions could not be interpreted.
As mentioned in Chapter 5.1.2, further interpretability work (for instance, token-level SHAP analysis or exemplar text inspection) would be required to determine which linguistic patterns are most influential.
We could also recalibrate of the predicted quantiles or Bayesian ensembling to better align empirical coverage with nominal confidence levels.

References

[Chen and Guestrin(2016)] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[Dunn and Smyth(2005)] Peter K. Dunn and Gordon K. Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280, 2005. doi: 10.1007/s11222-005-4070-y.

[Hoerl and Kennard(1970)] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[NCEI(2025)] NCEI. Noaa storm events database. `https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/`, 2025. Accessed: 2025-10-30.

[Tweedie(1984)] M. C. K. Tweedie. An index which distinguishes between some important exponential families. In J. K. Ghosh and J. Roy, editors, *Statistics: Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pages 579–604, Calcutta, India, 1984. Indian Statistical Institute.

A Appendix

Rank	Feature	Mean |SHAP| (Property)
1	EventNarr_emb319 (Narrative text feature)	1.541
2	Event Type: Hurricane (Typhoon)	0.495
3	Mean Nearby Damage (past events)	0.341
4	Direct Injuries	0.316
5	Duration (hours)	0.302
6	Mean Recent Damage (past events)	0.299
7	EventNarr_emb92 (Narrative text feature)	0.243
8	Event Span (km)	0.214
9	EventNarr_emb378 (Narrative text feature)	0.157
10	EpisodeNarr_emb96 (Episode text feature)	0.149
11	EventNarr_emb223 (Narrative text feature)	0.118
12	EventNarr_emb280 (Narrative text feature)	0.117
13	EventNarr_emb138 (Narrative text feature)	0.111
14	EventNarr_emb166 (Narrative text feature)	0.085
15	EventNarr_emb327 (Narrative text feature)	0.084
16	EpisodeNarr_emb319 (Episode text feature)	0.081
17	EpisodeNarr_emb39 (Episode text feature)	0.081
18	EventNarr_emb257 (Narrative text feature)	0.077
19	EpisodeNarr_emb117 (Episode text feature)	0.074
20	EventNarr_emb334 (Narrative text feature)	0.070

Table 2: Top 20 features by mean absolute SHAP value for the Property Damage model.

Rank	Feature	Mean |SHAP| (Crop)
1	Neighbor events (100km) (Count of nearby events)	0.801
2	Neighbor events (24h) (Count in last 24h)	0.686
3	Neighbor events (72h) (Count in last 72h)	0.685
4	Neighbor events (6h) (Count in last 6h)	0.679
5	Latitude (event location)	0.652
6	Location: Specific County/Area #1	0.612
7	Longitude (event location)	0.610
8	Location: Specific County/Area #2	0.588
9	Duration (hours)	0.573
10	Location: Specific County/Area #3	0.412
11	Event Span (km)	0.352
12	Mean Nearby Damage (past events)	0.300
13	Location: Specific County/Area #4	0.254
14	Event Type: Severe Ice/Cold event	0.159
15	EpisodeNarr_emb2 (Episode text feature)	0.114
16	Event Type: Drought	0.109
17	EpisodeNarr_emb266 (Episode text feature)	0.100
18	EpisodeNarr_emb263 (Episode text feature)	0.099
19	EpisodeNarr_emb97 (Episode text feature)	0.093
20	Event Type: Hurricane (Typhoon)	0.090

Table 3: Top 20 features by mean absolute SHAP value for the Crop Damage model.

B Electronic appendix

The code produced in this work can be found on: `https://github.com/ngabrielaxej/Stormevents_Damage_Prediction` The benchmark test was run on a personal computer with AMD Ryzen 9 6900HS with Radeon Graphics, 3.30 GHz, and 8 physical cores. It took 8 hours to train.