Property Damage Predictor

Greg Welliver

Problem Statement & Objective

Problem Statement & Objective

- Question: Is it possible to predict the amount of property damage that a storm will cause, given the characteristics of that storm?

- The objective of this project was to investigate the characteristics of storms that cause property damage. The end result is a predictive model where a storm can be input into the model, and the model will return how much property damage will be caused by the storm.

Data Overview

Data

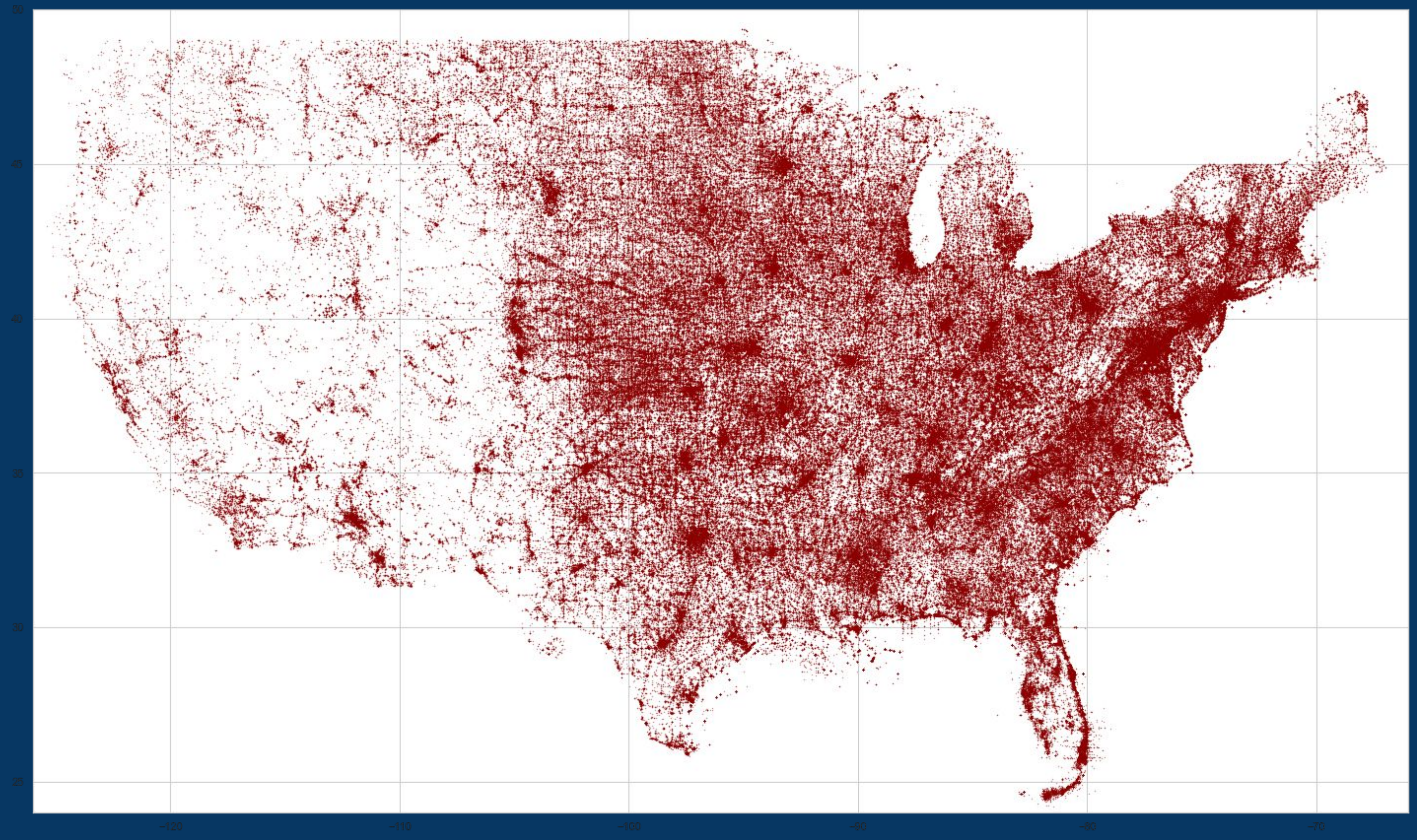
- Data sources: Storm detail from Iowa Environmental Mesonet, population information from census.gov, home price index from FHFA.gov

- Data highlights:
  - Collected information on 1,555,648 storms; from 2000 - 2022
  - The data includes a high number of outliers and high number of zeros, making this a fairly irregular dataset.

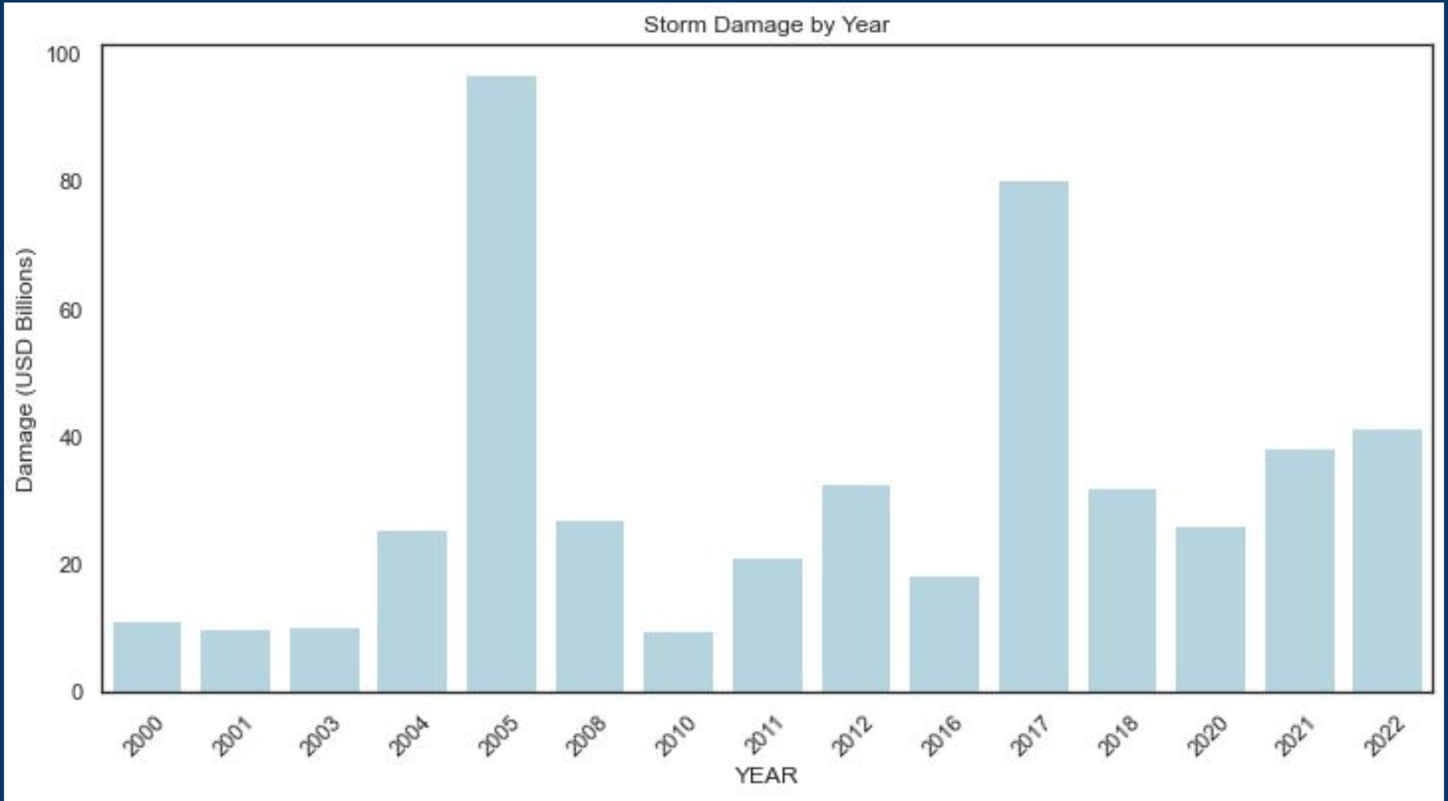
- Data for each storm included:
  - Storm ID, date, state, storm type, storm begin time, storm end time, direct & indirect injuries, crop damage, storm magnitude, flood cause, category, tornado scale, tornado size, storm latitude and longitude, event narrative

Exploratory Data Analysis Highlights

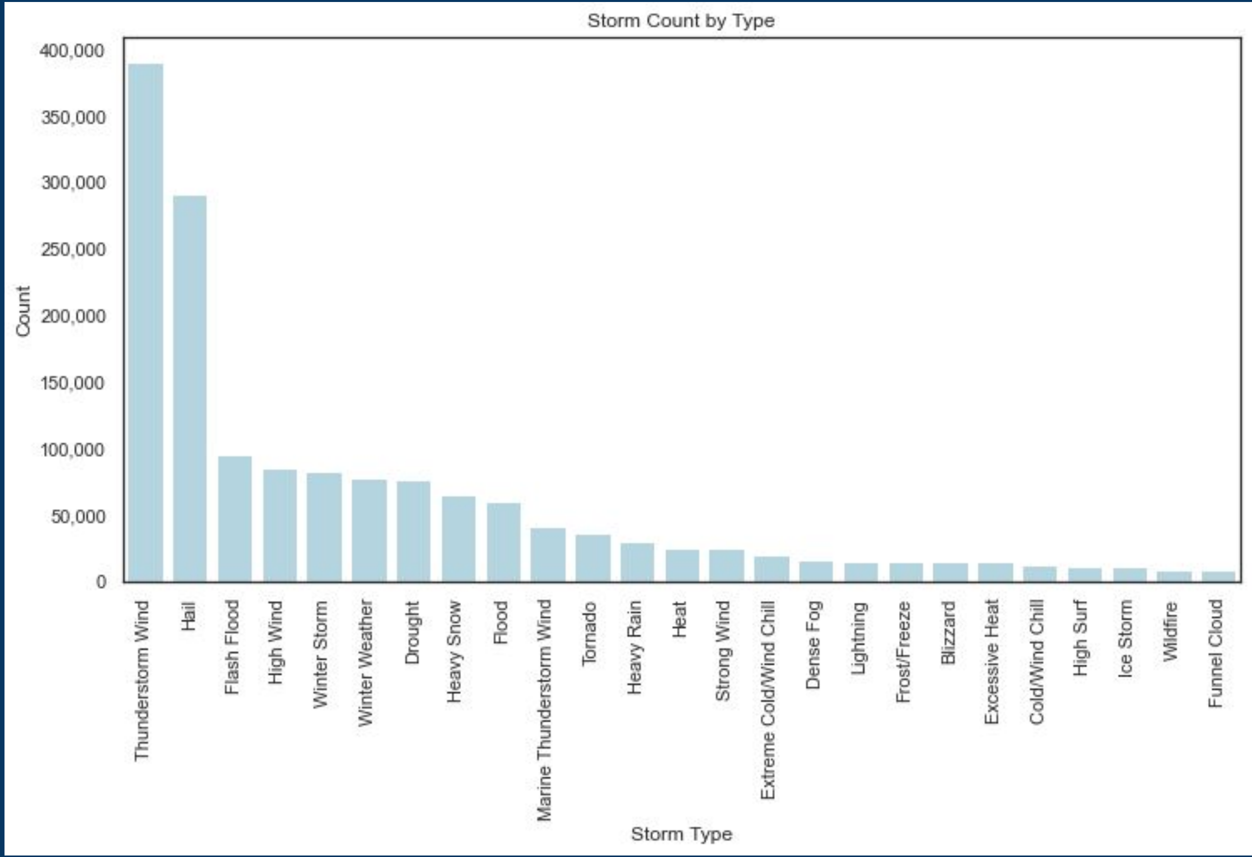
Geographic map of storms



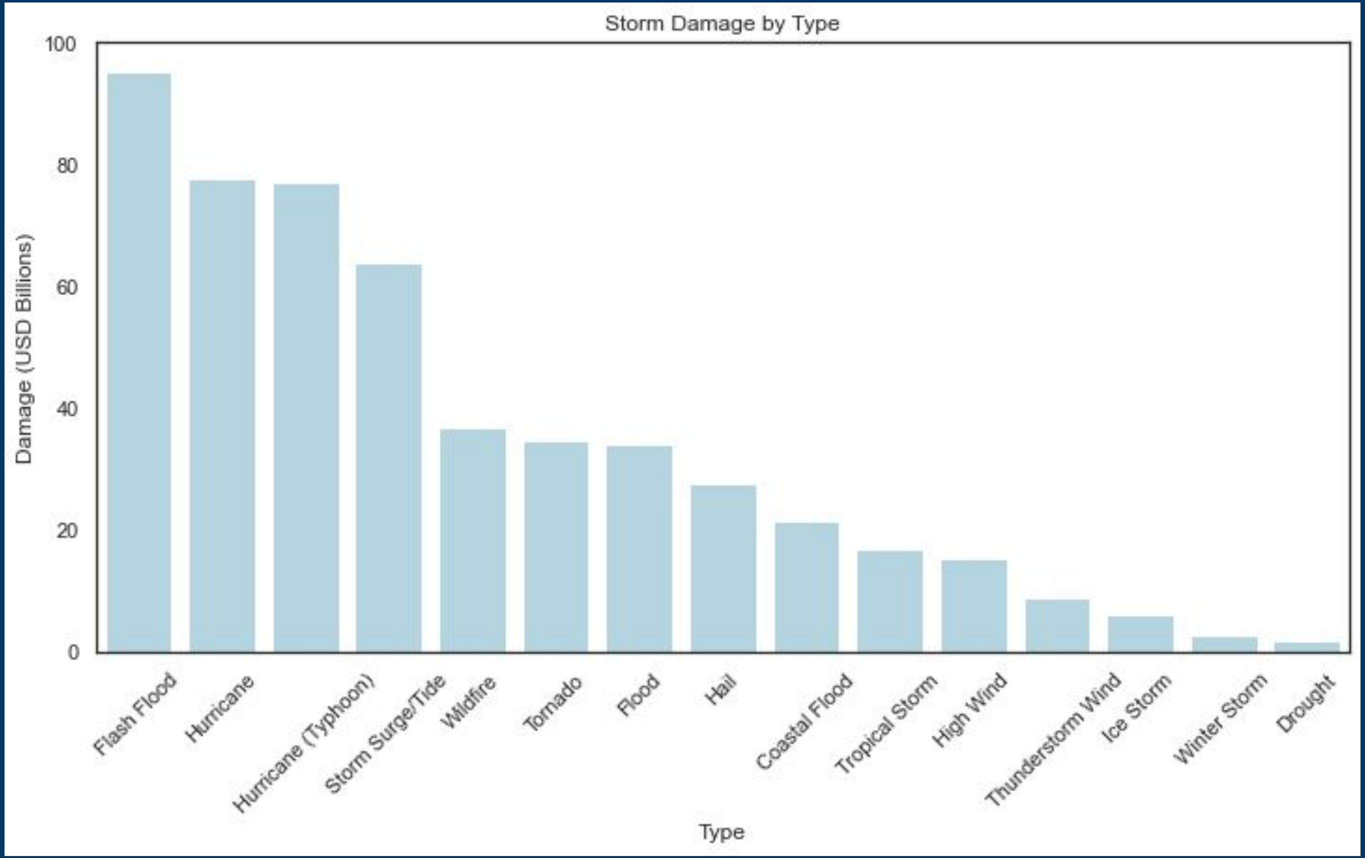
Storm damage by year



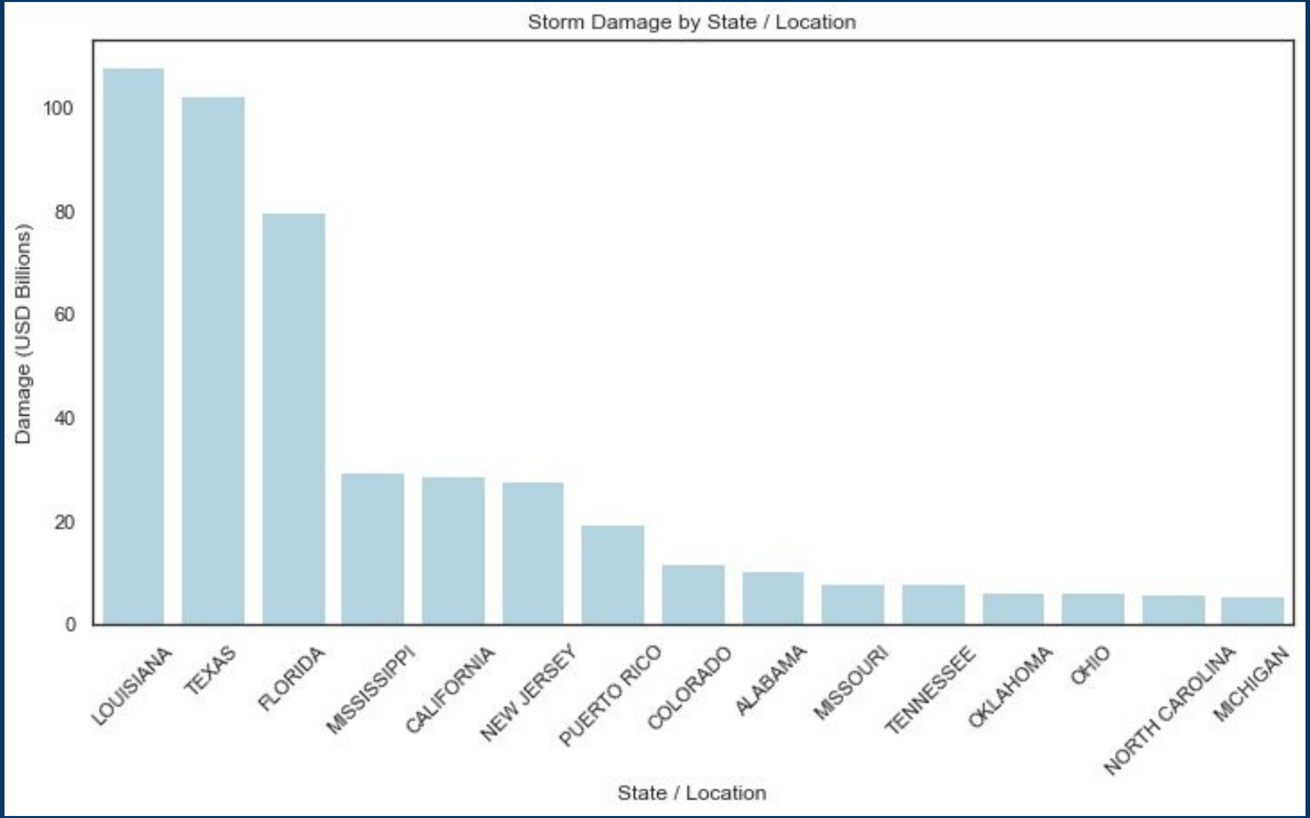
Storm count by type



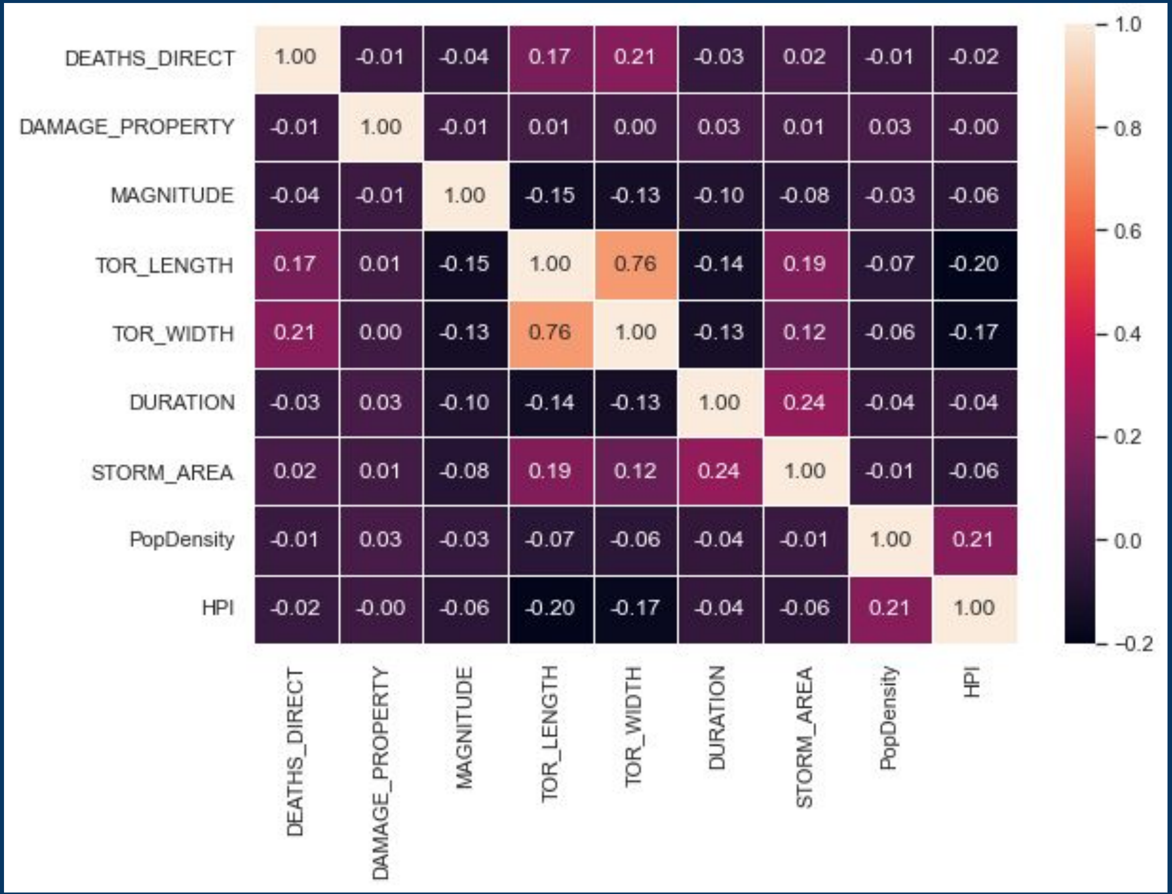
Storm damage by type



Storm damage by state / location



Correlation heatmap



Algorithms and Modeling

Model Selection

- I chose to implement the following algorithms in my pursuit of the most effective model for this regression problem.
  - Random Forest
  - LightGBM
  - XGBoost
  - Tweedie Regression
- Each of the algorithms were relatively close in performance. **LightGBM performed the best, followed by XGBoost, Random Forest, and Tweedie**

Metric Selection

- I looked at MAE, MSE, and RMSE as evaluation metrics for these models. These are appropriate for this type of regression problem.

- I was able to improve the model by over 70% from baseline to final. I achieved this by iterating on removing outliers, reducing features, and hyperparameter tuning.

Model Metrics

Algorithm	MAE		MSE		RMSE		MAE	
	Training	Test	Training	Test	Training	Test	DIFF	Comments
Random Forest								
Baseline	10,192	27,985	2,908,505,925	19,156,222,167	53,931	138,406	175%	Evidence of overfitting
Tuned	27,157	27,739	17,531,346,006	19,479,915,784	132,406	139,571	2%	Suitable
Light GBM								
Baseline	25,751	26,708	15,935,606,039	18,901,718,418	126,236	137,484	4%	Suitable
Tuned	25,862	26,861	16,013,959,656	18,950,807,358	126,546	137,662	4%	Suitable
XGBoost								
Baseline	25,001	26,917	14,684,962,256	19,102,140,608	121,182	138,211	8%	Suitable
Tuned	25,275	26,860	15,048,715,958	19,034,888,273	122,673	137,967	6%	Suitable
Tweedie Regression								
Baseline	29,731	30,585	18,761,667,052	20,554,164,254	136,973	143,367	3%	Suitable
Tuned	28,632	29,326	21,214,757,723	22,392,103,136	145,653	149,640	2%	Suitable; power = 1

Future Improvements, Opportunities for Pursuit

Future Improvements

- It would be interesting to attempt to separate the impact of increasing property prices and storm severity on total damage. With property prices increasing substantially in the United States over the past few decades, this would have a meaningful influence.

- Explore additional datasets that could provide predictive value