Predicting Flood Damage to Buildings

Based on FEMA Flood Insurance Policy Claim Data
Springboard Capstone Two, Spring 2020

Context

Flooding poses a risk to human health and causes significant damage to buildings and assets across the world every year. In the US, flooding impacts residents of all 50 states, particularly those living in coastal communities. As experienced just this month with Hurricane Laura in Texas and Louisiana, flooding can displace families, destroy homes, and result in dangerous living conditions for entire communities. The frequency and severity of flood events associated with extreme storms and hurricanes continues to intensify as the effects of global climate change evolve.

Question

Can we use building and site characteristics to help homeowners and property developers predict and possibly prevent severe flood damage?

In particular, can a model that predicts flood damage in dollars be useful in making the case for increasing building resilience through strategies such as elevating the first floor?

Data Source

The Federal Emergency Management Agency (FEMA) manages flood insurance programs in the United States. Any property located within a FEMA-designated flood zone is required to purchase flood insurance from FEMA. As such, FEMA has collected many years-worth of data describing flood policies and flood claims.

This project uses the FEMA Flood Insurance Claim dataset released for the first time by FEMA in 2019. The dataset includes information regarding building and site characteristics for properties that have submitted a flood insurance claim between the 1970s and today. Please refer to the data dictionary for more detailed information on the dataset variables.

Data Wrangling

<u>Missing Values</u>
This dataset includes over 2 million flood policy claims from as far back as the 1970s. As such, it includes many missing and invalid values. To maintain as much information as possible, no missing values were removed. Instead, new indicator features were created to capture where values were missing, then those missing values were filled

with either the median or mode of the column, based on the nature of the variable. The data dictionary includes a complete summary of how missing values were handled for each variable.

Flood Zone

FEMA's flood zone designations have evolved over time. As such, the flood zone variable contained references to outdated zones and multiple zones which are now classified together. To simplify and make this variable relevant for today, a new flood zone variable was created based on the primary zone letter (A, AE, C, V, VE, X, and 'other').

Date Variables

To extract more information from the construction date and the date of the flood event, new features were also created to represent the year and month of those events.

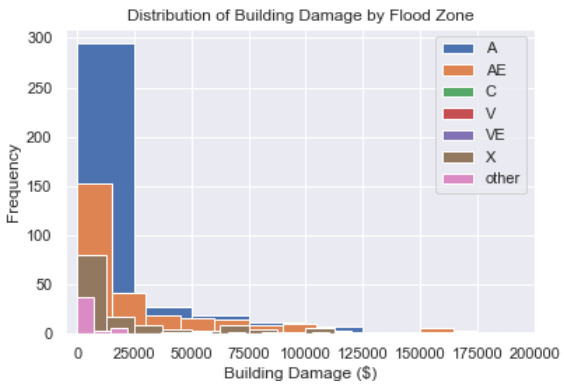
After wrangling and cleaning this dataset and creating dummy features out of the categorical variables, the final dataset shape was 2,418,007 rows and 203 features.

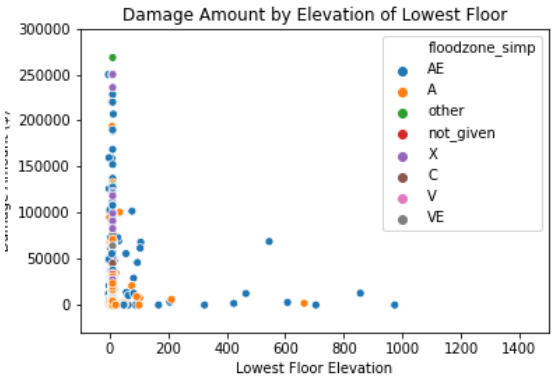
Exploratory Data Analysis

Flood Zone

Most flood damage claims occur in the A and AE flood zones, these represent the 100-year flood zone which tend to be larger in area than other flood zones, but are also closer to the coast or bodies of water than the X zone or the 500-year flood zone. However, the range of the amount paid on the building damage is similar among the 100 and 500-year zones.

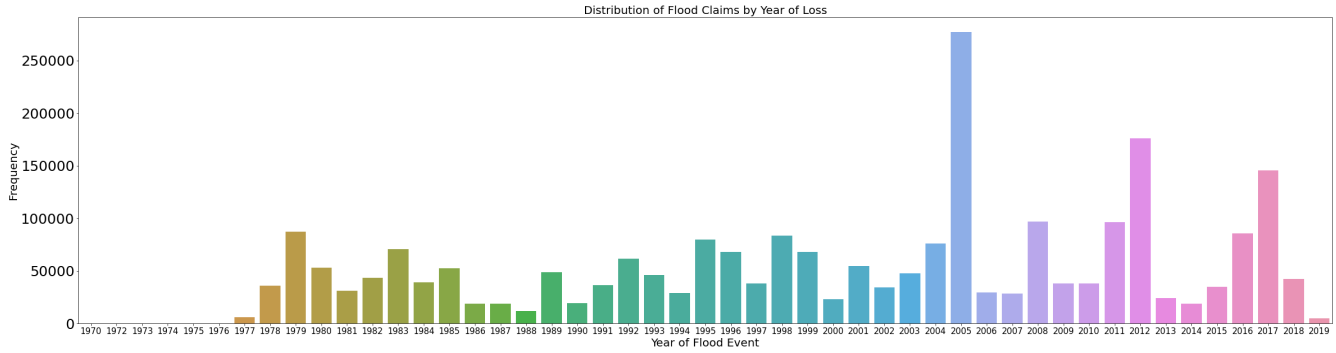
Additionally, buildings in the 100-year zone appear to be more likely to have elevated first floors. This characteristic is likely because buildings in the 100 year flood zone has a higher risk of flooding than the 500-year zone.





Date of Loss

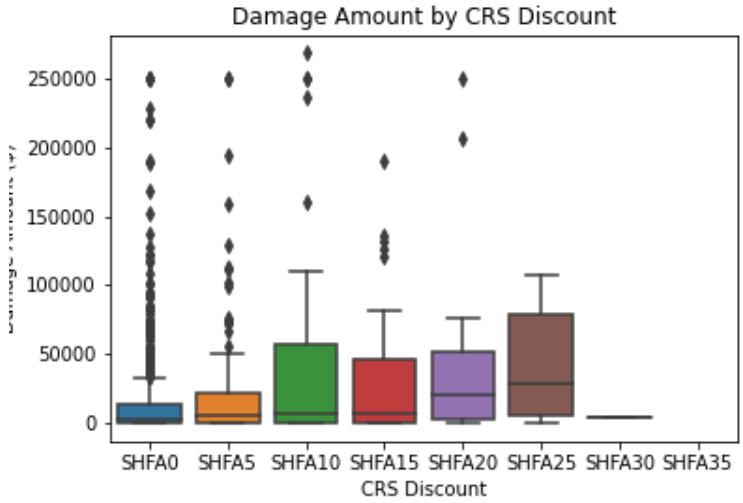
There is significant variation in the number of claims submitted each year. However, it is clear the overall trend is increasing with significant flooding years occurring in 2005 (Hurricane Katrina), 2012 (Super Storm Sandy), and 2017 (Hurricane Harvey).



Community Rating System Discount

The Community Rating System (CRS) is a voluntary FEMA program that awards a discount on flood insurance premium to communities that employ flood hazard mitigation strategies on a community scale. The more floodzone management practices in place, the higher the discount on insurance premiums. Interestingly, it appears that communities with a higher CRS score (SHFA25) tend to pay more for building damage than communities with lower CRS scores (SHFA0).

However, this characteristic may be driven by fewer communities participating in the program as well as communities at higher risk having more incentive to employ more floodzone management strategies.



Model Selection

Both linear regression and tree-based algorithms were tested to create a model of the FEMA flood damage claim dataset. Metrics including RMSE and R2 were used to evaluate the performance of each model tried.

Linear Models

Linear regression was used as a quick initial model to fit. Using this initial model, residual plots were then used to complete additional feature engineering on several numerical features to allow the model to further understand this complex dataset. While these linear models showed high error and low R2 values, the supplemental feature engineering did improve results significantly. Including regularization with a Ridge model did not meaningfully improve model metrics.

Tree-Based Models

A decision tree regressor showed a significant improvement over the linear models and improved the RMSE metric by more than $10,000, making its predictive performance more appealing. Moving from a decision tree to an ensemble method with a random forest regressor proved to be the most effective model with the lowest RMSE and highest R2 value.

	Model	RMSE	MSE	MAE	R2
0	Initial Linear Regression	51637.851494	2.666468e+09	24587.314699	0.043978
1	Linear Regression_New Features	51509.955715	2.653276e+09	24575.968570	0.048708
2	Ridge Regression_alpha 0.1	51509.922542	2.653272e+09	24575.948201	0.048710
3	Ridge Regression_alpha 1.0	51509.674187	2.653247e+09	24575.801688	0.048719
4	Ridge Regression_alpha 10.0	51506.687010	2.652939e+09	24575.023541	0.048829
5	Decision Tree_no max depth	50940.682618	2.594953e+09	20628.905221	0.069619
6	Decision Tree_max depth 20	44352.663573	1.967159e+09	17738.136575	0.294705
7	Decision Tree_max depth 5	47845.999890	2.289240e+09	22923.252008	0.179228
8	Random Forest_estimators 10	40119.804842	1.609599e+09	17103.133113	0.422903
9	Random Forest_estimators 50	38946.508137	1.516830e+09	16968.584523	0.456163
10	Random Forest_estimators 100	39139.883893	1.531931e+09	16966.830714	0.450750

Model Results

The random forest regressor proved to be the best model for this dataset. With this model, a user may be able to predict future flood damage costs based on simple building and property information.

The features with the most influence over building damage include the total amount of coverage of the insurance policy, the month of the flood event, and the year of construction of the building. Base flood elevation and the elevation of the lowest floor are also within the top 10 important features suggesting some design considerations around building elevation may be a useful factor in reducing potential damage costs.

Feature	Importance
totalbuildinginsurancecoverage	0.1639693
monthofloss	0.0610176
constructionyear	0.0598852
policycount	0.0433211
basefloodelevation	0.0350354
lowestfloorelevation	0.034826
nbyear	0.0341901
state_MS	0.0215496

Limitations and Next Steps

This dataset includes flood claim data for properties that hold flood insurance through FEMA. Flood insurance policies are only issued to properties that exist in a flood plain, therefore this sample represents properties most at risk to flood events. However, it does not capture all flood damage in the US. As sea levels rise and weather patterns intensify, FEMA's flood mapping efforts are struggling to stay updated with current conditions and especially with future conditions. Similarly, this dataset does not include properties in flood zones which have not experienced flooding over the last 50 years. This represents a gap in data coverage and may result in biased results.

This dataset also includes many datapoints with questionable interpretation including very large and very small (negative) elevation values. This may indicate inconsistent units or changes database management standards that have evolved since the 1970s.

Further investigation of the invalid or unexpected values may help improve interpretation of the dataset and model results. Additionally, supplemental datasets such as data describing weather patterns or other flooding events, may help reduce any bias associated with the FEMA data limitations.