SIADS 696 Milestone II Project Report – Team 4

Predicting Housing Damage From Hurricanes

Scott Powell, Kingsley Reeves, Jr., and Joe Schweiss

Hurricanes provide a mix of rain, wind, and storm surge flooding that threaten both property and people. According to a 2024 study from the National Oceanic and Atmospheric Association (NOAA), climate scientists note that while the number of hurricanes has not increased since 1979, the number of major hurricanes increased over the last 20 years[1]. Current projections point to an increase in active Atlantic hurricane seasons from now through 2049, increasing both inactive and active seasons "at the expense of a reduction in the near-normal seasons[2]." This variability will lead to further challenges in the already difficult field of disaster preparedness. Innovations must be made in hurricane response efforts as increasing hurricane intensity shows no sign of abating.

This project aims to develop a machine learning model that can predict the number of housing units at the county level along a storm's path that will experience severe damage after a hurricane. Hurricanes are the most damaging natural disasters that frequently occur in the United States. Delayed and inadequately resourced responses put people at risk, increase time for a community to recover, and cause major political fallout. Developing the ability to predict severe damage can help with disaster planning and response. For instance, if many homes are in danger of severe damage, the Red Cross can arrange for more shelter space and supplies. The US Department of Housing and Urban Development (HUD) can conduct analysis on how many temporary vouchers they may need to issue. The Federal Emergency Management Agency (FEMA) can anticipate the additional supplies or resources that may be needed at the state level, and can issue mission assignments more effectively during response operations. With better predictions, the entire emergency management community can more effectively help people during their greatest time of need.

For this project we compared various supervised models, and employed unsupervised methods for feature creation. Data for this project comes from the FEMA Individual and Households Program, FEMA disaster declarations, NOAA, and the Census Bureau. The unsupervised methods employed attempted to categorize counties both by housing characteristics and economic characteristics. After applying principal component analysis, the counties were clustered with both agglomerative clustering and k-means clustering. The resulting clusters were compared with Davies-Bouldin and Silhouette scores, and the optimal clusters were then added to the supervised models as features. After inclusion in the supervised model, it was determined the clusters did not have much predictive power in determining hurricane damage.

In our supervised learning approach, we compared models from a variety of families and performed hyperparameter tuning using RandomSearchCV and 10-fold cross validation in order to determine the best option for correctly predicting the number of housing units experiencing major storm damage at the county level. The RandomForestRegressor model from Scikit-Learn was ultimately selected for deeper analysis and predictive error evaluation in hopes to make future improvements. Our main finding from the model is that it tends to overestimate damage more often than it underestimates, but the underestimates are further away from the target on average.

Related Work

The study by Ma et al. examines the impact of housing tenure and income levels on home damage due to Hurricane Maria in Puerto Rico. Similar to our approach, the authors leverage available FEMA data. Unlike our

[1] Hosmay Lopez et al. in References
[2] AOML Communications in References

study, the authors only consider a single location (i.e., Puerto Rico), the impact of a single storm (i.e., Hurricane Maria), and do not include hurricane strength data or primary Census Bureau data. Further, the study does not use machine learning techniques to generate predictions.

More similar to our project, the study by Klepac et al. employs machine learning techniques to enable better hurricane preparedness in communities. The authors consider several machine learning models, including k-nearest neighbors, decision trees, random forests, and gradient boosting trees. Among the model features considered are building, hazard, and geospatial data. In contrast to our approach, the authors only include data from four hurricanes in coastal locations in their dataset whereas we track storm paths and resultant damage in all areas of nine US states and Puerto Rico at the county level.

Pinelli et al. also developed a prediction model to estimate expected damage to residential buildings due to hurricanes. The probabilistic model developed by the authors is based on laboratory studies, surveys of after-storm damage, insurance claims data, and engineering assessments. Unlike our approach, the authors use Monte Carlo simulation to determine their predictions. In further contrast to our modeling approach, the authors of this paper are merely offering a proposed modeling framework; their model is not being employed and tested, and no real data is being used. Thus, their model is not assessed for accuracy.

Pilkington and Mahoud developed a predictive model to predict storm damage. However, the model these authors employ is not limited to the consideration of wind damage but also includes damage resulting from precipitation. Further, the authors focus more generally on economic damage to an impacted region versus our more narrowly focused attention to damage to residential housing. The authors propose a neural network model with inputs that include population, landfall location, wind speed, pressure, storm surge, and precipitation. While we include some of these same features in our model, the feature set proposed in this study is more expansive because of the authors' consideration of multiple hazards. Moreover, in contrast to this study, our damage estimate is more fine tuned and is at the county level versus a higher aggregate level.

Li and Gu also present a model to predict building damage following a hurricane that employs a neural network. In contrast, we explore the use of supervised learning models like random forests, decision trees, and gradient boosting. Further, unlike our model that uses storm characteristics and housing data for the regions of interest to predict damage, the model presented in this paper processes image data as its input. The authors use a convolutional neural network model to analyze post hurricane satellite images to detect damaged buildings.

Our project is an extension of the Milestone I project that Scott Powell worked on. In that project, he and his collaborators - Qunkun Ma and Jerry Sweitzer - used the FEMA disaster declarations dataset and NOAA data to examine correlations between disaster frequency and climate change. The current project goes beyond looking for a correlation between climate and disaster frequency and instead tries to predict the number of damaged houses after hurricanes.

Data Sources

OpenFEMA datasets (OpenFEMA Data Sets | FEMA.gov)

- Disaster Declarations Summaries V2: This dataset is downloaded as a CSV file and contains all federal disaster declarations dating back to January 1953. This dataset provides many variables, a few of which were useful for our analysis. Since our level of analysis is the county, the FEMA dataset helped us isolate which counties were impacted by storms as the declarations are recorded by county. It also helped us separate each impacted county by declaration number, storm name and year. This was crucial for helping us organize our data from the Census Bureau and NOAA.
- Housing Assistance Program Data - Owners V2: Our target variable came from this dataset. For each disaster, aggregated information is provided by county for the total number of valid registrations.

Registrations are categorized by amount of assessed damage. For our purposes, major damage is greater than $20,000. This dataset covers disasters since November 2002, and is returned as a CSV file.

- Housing Assistance Program Data - Renters V2: In addition to homeowners, we included damage to rented dwellings. Unlike the owners dataset, the renters dataset does not aggregate by dollar amount, but by damage category. In this case, we employed the 'major' damage rating for our target variable. This number is combined with the total number of owned dwellings that have greater than $20,000 damage. This dataset covers disasters since November 2002. This dataset is downloaded as a CSV file.
- Individuals and Households Programs - Valid Registrations V2: This dataset, returned as a CSV, was ultimately not used. It contains individual registrations by disaster since November 2002. This is a very large - 10 GB - dataset as it does not contain aggregated data. We downloaded the data, converted it to a parquet file, then grouped the registrations by disaster number and county. However, after completing this, we found that we did not gain more information than we had from the aggregated datasets. In interest of computing power and storage of our final product, we opted not to include this dataset.

NOAA HURDAT2 (Index of /data/hurdat)

- The NOAA Hurricane database (HURDAT2) covers data for recorded hurricanes in the Atlantic basin from 1851-2024, and is downloaded as a TXT file. For our purposes, we restricted the analysis to hurricanes from 2003-2024 to match the FEMA registration data. Storms are recorded by name and date, which we used to merge with the other datasets. From this dataset, we obtained the latitude and longitude of landfall, wind speed at landfall, and minimum central pressure at landfall. Wind Speed and central pressure served as a measure of the storm's power. The latitude and longitude helped calculate the distance of each county from landfall, accounting for decreasing storm power based on distance.

Census Bureau

- American Community Survey 5-year estimates (Census Bureau Data): The American Community Survey provided data at the county level for the majority of the 2003-2024 timeframe. Data was downloaded in CSV files for each relevant time period. Specifically, we employed tables DP04 for housing data and CP03 for economic data per household. The DP04 table provided the number of houses, number of occupied houses, total number of houses by decade of construction, and total number of houses by type. The CP03 table provided income level by household and median household income for unsupervised learning. The 2000 Decennial Census, which is aggregated at the state level, was used to interpolate missing values prior to 2005. The aggregated decennial data was redistributed across counties using county estimates obtained from the 2005 5-year estimate.

- Centers of Population (Centers of Population): The centers of population provided a latitude longitude for the center of population for each county. Data is provided on the webpage and was simply copied into a TXT file. This was used to calculate the distance from landfall for that county. Puerto Rico is not included in this set, so the geographic center of Puerto Rico as provided by Google Maps was used.

Feature Engineering

A detailed breakdown of all features used for supervised learning can be found in the Features for Supervised Learning Appendix. The summary below provides the high-level features, some of which were further modified during model training and evaluation.

Supervised Learning Features:

- Total damaged houses: This feature was our target feature for analysis. In order to obtain this feature, we combined numbers from the FEMA Housing Assistance Program for both renters and owners. This feature

required very little engineering as it combined three numbers: major damage for renters, damage from $20k to $30k, and damage greater than $30k.

- Max Wind: Represents the maximum wind speed for a hurricane at first landfall in the United States. This data came from the HURDAT2 dataset. In order to obtain this data, we had to filter according to storm status ('HU') and landfall ('L'). Since some hurricanes made landfall in Cuba, Mexico, the Bahamas, or the Windward/Leeward Islands, we had to filter for landfalls by latitude and longitude. With the southernmost point in the continental United States being the Florida Keys, we filtered for landfalls above 24.N. This could potentially include landfalls in the Mexican state of Tamaulipas. Storms making landfall in this area would likely impact Texas, so this was not deemed a concern. In order to ensure Puerto Rico was included in the dataset, we had to place a 'box' around Puerto Rico using latitude and longitude. Similarly, we had to designate an area around the Bahamas to exclude landfalls in the Bahamas since at least one storm transited the Bahamas before impacting Florida. We did not similarly place a box around the US Virgin Islands as that particular territory was not a top ten location for hurricane landfalls.
- Min Pressure: Represents the central pressure of the hurricane at landfall, and was obtained with the same methods as the max wind feature.
- Distance from landfall: Serves as a proxy for the weakening of a storm the further inland it moves, or for the weaker outer edges of a storm. This variable was calculated with the centers of population coordinates and the latitude longitude coordinates. We used geopandas distance calculation with the latitude and longitude for the county centroid and hurricane landfall as tuples for the input. The output of the function was distance from landfall in miles.

- Closest Pass: After working with our initial model, we determined that we needed more refined information regarding storm strength. Our initial model showed that distance from landfall had the strongest correlation with the number of damaged houses. Using the HURDAT2 dataset, we modified our code to select all occurrences for a storm where it was at or above tropical depression strength. Similar to the Max Wind/Min Pressure, we filtered based on latitude and longitude to get storms that impacted the US. Using the population centers obtained from the Census Bureau, we used geopandas to calculate the 'closest pass' a storm made to a county. We then selected the wind speed, max pressure, and calculated the distance during that pass. All three of those parameters became features in our supervised learning model.

Unsupervised Learning Features for Clustering:

- Census DP04 - Selected Housing Characteristics: This table from the ACS 5-year estimates (2023) provided information regarding the number of houses built by decade, and numbers of structure by type for each county. This specific set was chosen as it encompassed all decades for building construction, a variable that was also employed in supervised learning. The houses by structure could further enhance analysis as an urban county would likely have more multi-unit structures, and a rural county would likely have more single unit structures. Additionally, some counties would be more likely to have larger numbers of mobile homes, people living in vans, houseboats or other structures highly susceptible to storm damage. This was used to create clusters that could account for local construction techniques.
- Census CP03 - Selected Economic Characteristics: This table also from the ACS 5-year estimates provided information by county regarding number of houses at specified income levels. It also provided median household income by county. We employed this information to cluster counties economically, with the assumption that more prosperous counties might have better construction and possibly fewer disaster registrations. The lower number of registrations might indicate higher rates of fully insured homes.

Part A – Supervised Learning

The data was divided into X and y components; the features and target, respectively. These were split into training and testing sets, using a ratio of 80% training to 20% test. The X training set was then used to fit a column transformer pipeline, consisting of StandardScaler for the quantitative features and OneHotEncoder for the categorical features. The X training and test sets were then transformed according to this fit. The y train and y test sets were transformed separately using Numpy to convert to natural log plus one in order to correct skewness in the target values. All of these actions were done in this manner to avoid data leakage. Stratification was used to ensure that each split would receive a distribution of values from the storm category, which we hoped would spread non-zero target values evenly.

Representatives from multiple model types were selected for an initial performance test, including linear, ensemble, tree-based, and neural-networks. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 scores were used as metrics. MAE was ultimately selected for its ability to minimize the impact of outliers and for easy interpretation. 10-fold cross validation over the training set was used to conduct multiple tests over different datasets that could then be averaged together for greater accuracy and to maintain the integrity of the test set. The models that stood out as the best options based on mean MAE scores (Table 1) included:

- RandomForestRegressor (RFR), an ensemble tree-based model from Scikit-Learn that combines predictions from multiple trees by averaging their individual predictions;
- HistGradientBoostingRegressor (HGBR), an ensemble tree-based model from Scikit-Learn that combines predictions sequentially, allowing each subsequent tree to refine the previous prediction; and
- MLPRegressor (MLPR), a neural network model from Scikit-Learn using multi-layer Perceptron and stochastic gradient descent.

Hyperparameter tuning was conducted on the above models using RandomSearchCV with 10-fold cross validation over the training dataset. After tuning, the models achieved the following scores:

Table 1	**RFR**	**HGBR**	**MLPR**
Mean MAE with s.d.	1.004 +/- 0.095	0.979 +/- 0.071	1.046 +/- 0.084

While HGBR would be the slightly better choice, the model does not provide a feature_importance_ attribute. For the sake of more complete analysis, RFR was selected for the remainder of our work instead.

Supervised Learning Evaluation

The difference between our target values and predicted values were measured. The RFR model predicted within 10 damaged housing units of the target 61% of the time, within 50 housing units 81% of the time, and within 500 housing units 98% of the time. The four largest errors were underpredictions, with the worst error falling 3989 housing units short of the target. This range of error does frequently avoid overestimating the impact of storms, which would result in sending too many resources to an area that does not need them, but the level of underestimation would likely result in too few resources going out to areas of greater need. RFR was selected for more detailed analysis. Despite the shortcomings, this model outperformed the others. Histograms for prediction error are shown in Figures 1 and 2.

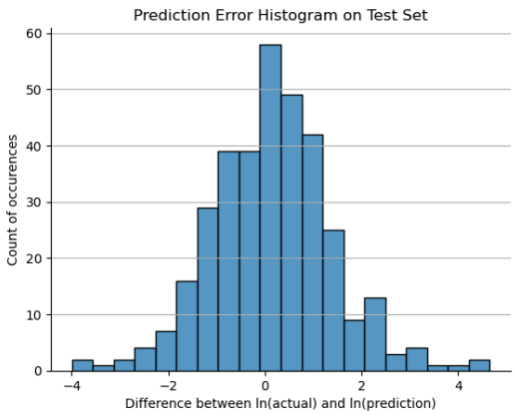


Fig 1. - Prediction Error Histogram

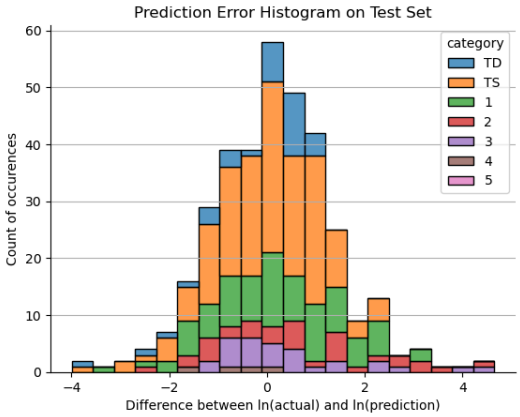
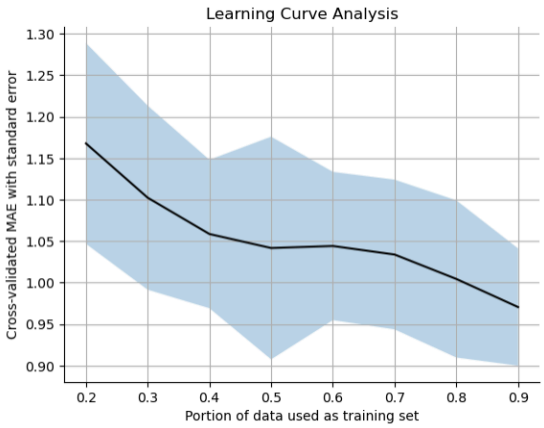


Fig 2. - Prediction Error Histogram by storm category

Learning Curve Analysis

We chose to aggregate our data sources into a dataset that contained information from multiple hurricanes spanning multiple decades and their impact on individual counties. This reduced thousands of rows to just 1,730 rows. This means one row per county per storm as observations. We wanted to understand if this amount of data was enough for the goals of our project, so we conducted a Learning Curve Analysis to study the impact of training data sample size on the model's performance, in Figure 3 at left.



The full dataset was broken into portions of the total dataset to be used for the training set, from 20% as the baseline, adding an additional 10% of the total dataset until 90% of the total was used as the training set. The MAE values were calculated using a 10-fold cross validation at each training set size, then the mean and standard deviation of the MAE was plotted for analysis. We found that:

1. There was a brief plateau between 50% and 70% training set size, after which the MAE and standard deviations proceeded to drop.
2. The change between 70% and 90% training set size is less steep than the change between 20% and 40%, but the standard deviation value is smallest at 90%, indicating less variation in the MAE measured during 10-fold cross validation and therefore more predictable outcomes.
3. This information points to the amount of data that was collected being insufficient to properly train the model and test its performance. More time should be spent in gathering additional sample data or engineering more features given the data that was collected.

Feature Analysis (Including SHAP and Feature Ablation)

We used Shapley Additive Explanations (SHAP), an explainer of machine learning model outputs that uses game theory to understand predictions, to explain the behavior of our model. We created a beeswarm plot indicating SHAP values, the impact on model predictions, for the most important features as determined by the SHAP explainer. Results are depicted in Figure 4.

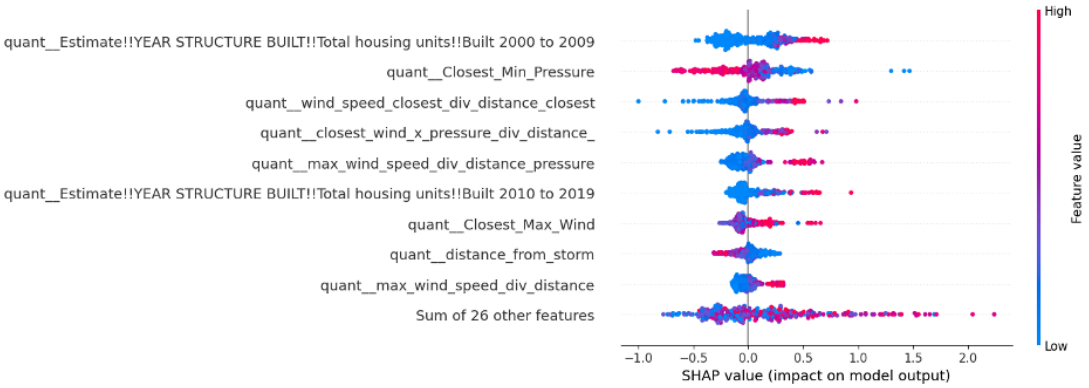


Figure 4 - SHAP Analysis

The most important features that provided consistent output were housing units built between 2000 and 2009 and the closest minimum pressure, followed by a mix of our features that we mathematically combined to create new features. We can see that the 26 features combined in the last part of the plot do not provide a consistent impact on the output based on the value of the features, meaning that we likely have a variety of features that are detrimental to the performance of our model. None of our categorical features are present in the top nine features. These top features are similar to our findings in a Permutation Importances test, but are not identical. Results of this analysis are displayed in Figure 5.

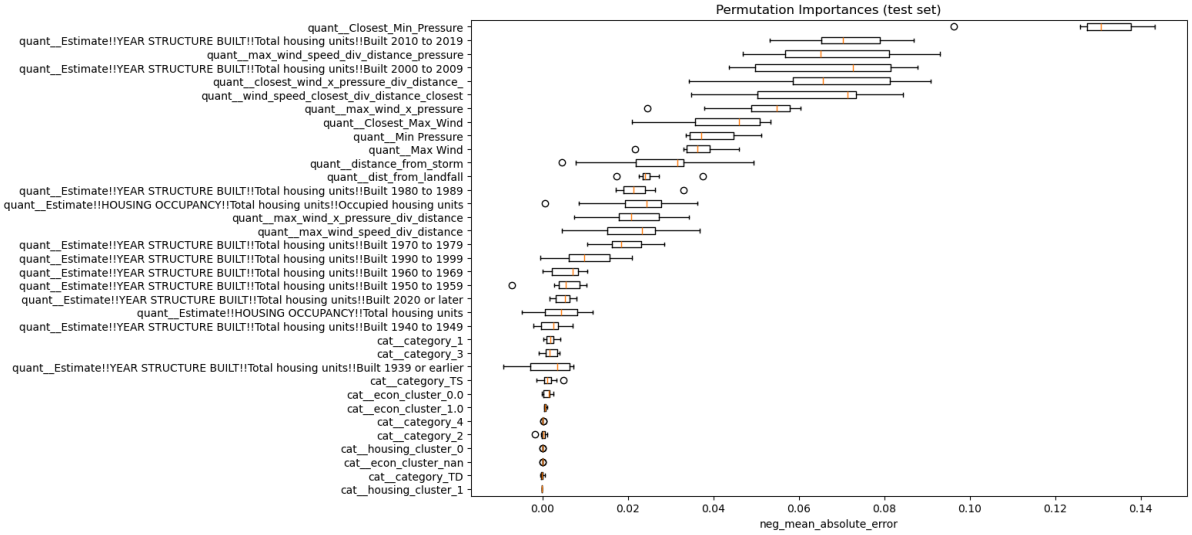


FIgure 5 - Permutation Importance

We conducted Feature Ablation Analysis by dropping groups of related features to see their impact on MAE, with results captured by 10-fold cross validation on the training set. For the sake of brevity, not all features will be named below.

Table 2 - Ablation Analysis

Feature Group	**Features Dropped**	**MAE**
No Ablation	None	1.005 +/- 0.095

Occupancy	Total Housing Units, Occupied Housing Units	1.002 +/- 0.097
Building Info	All Decade built features	0.996 +/- 0.084
Storm Category	Category 1, 2, 3, 4, TS and TD	1.007 +/- 0.099
Unsupervised Clusters	Housing and Econ Clusters	1.001 +/- 0.098

We found that:

1. Dropping the **Building Info** features did not result in a large increase in MAE, even though a few of these features were deemed important during SHAP analysis, but rather improved our MAE results. It may stand that some of these features are less correlated to the target value and may be detrimental to the model's performance. Relying only on **Building Info** and dropping the **Housing Units** information also improved the MAE slightly, but to a lesser extent. It would be worth selectively dropping some **Building Info** features in the future to find the best combination of housing unit and building information features to retain.
2. The original **Storm Category** information is somewhat useful in predicting damage to housing units along the path of the storm. This indicates that the way a storm changes along its final path is not entirely dependent on its strength at landfall, but this information has a minimal amount of predictive value.
3. The features discovered during the Unsupervised Learning portion of the project did not improve model outcomes. It would be worth repeating Feature Ablation with these categories dropped separately.

Sensitivity Analysis

The RFR model has a large number of parameters that can be set in order to adjust its accuracy and behavior. These are settings within the model that impact its output that are not learned from the data, and often the values that provide the most desired outcomes are not easily intuited. Understanding how sensitive a model is to hyperparameter adjustments can improve interpretability of a complex model, help increase robustness of the model, and gain insight into the potential limitations of the model.

We tested the sensitivity of multiple parameters and found **max_depth** and **max_features** to have the largest impact on MAE. 10-fold cross validation was used to calculate MAE over a range of values for both parameters mentioned while holding all others constant, both separately and in combination to explore the impact that these two parameters have on each other. We found that:

1. **max_depth,** the parameter setting the maximum depth of each tree in the forest, was the most sensitive in our model, though it reached a stable level of MAE after reaching a value of 15. The major plateau was at a value of 10, before which MAE values ranged from 1.466 to 1.010. After a value of 10 the results are much less varied, settling to a MAE close to 1.000. A value of 15 was selected as a best stopping point for having a low MAE, minimal standard deviation change, and little to no gains to be found in higher values. Increasing **max_depth** much more would likely result in a model that is overfitting training data.
2. **max_features,** the parameter that controls the number of features to consider when looking for the best split at leaf nodes, was the next most sensitive, but not nearly as sensitive as **max_depth.** The area with the widest range of MAE scores came between values of 1 and 7, with MAE value from 1.064 to 1.011. Values greater than 7 resulted in MAE scores between 1.019 to 1.004, but this was not very consistent. This leads us to believe that we have a minimal number of features that are consistently correlated to the target value.
3. **max_depth** and **max_features** were adjusted simultaneously to determine the impact they might have on each other. This turned out to be dominated by depth and performed very similarly to how they did when

adjusted separately. The mean MAE is highest at low values of **max_depth** and **max_features** and drops until **max_depth** reaches 10, then MAE settles near 1.004 with small changes in value over changes in the two parameters. There are no values where the MAE drops to zero. This lack of variation does indicate that our model does perform as well as can be expected with the modified settings and would be at higher risk of overfitting if the default values for both, no maximum depth or feature limit, was left in place. Results are depicted in Figures 6 and 7.

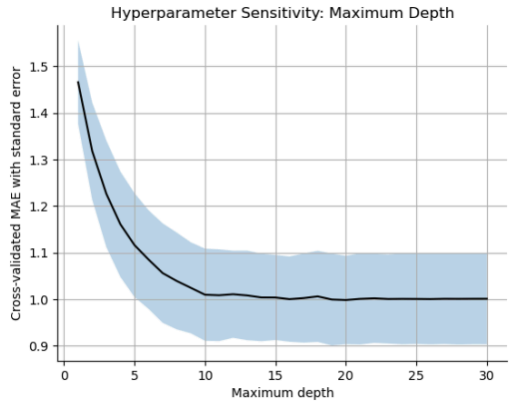


Figure 6 - Maximum Depth Sensitivity

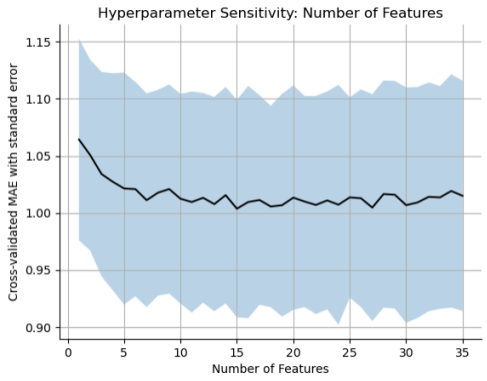


Figure 7 - Number of Features Sensitivity

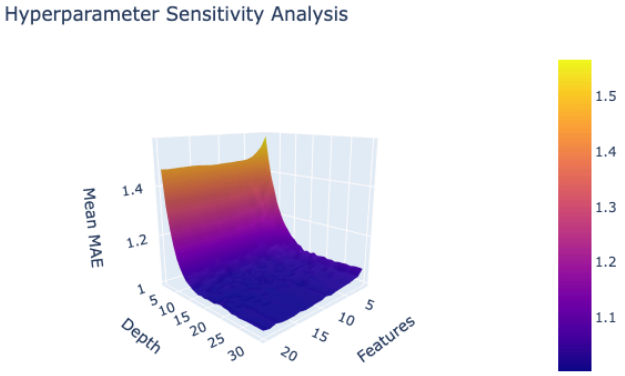


Figure 8 - Hyperparameter Sensitivity Analysis

Failure Analysis

In order to understand the shortcomings of our model, we performed an in-depth error analysis to find areas for improvement. We started with a visualization of the differences in our predictions versus the targets in the test set, binned by date. Note: these differences are actual differences, not the natural log of difference. The model overpredicted the target 192 times and underestimated it 154 times, but our overpredictions were off by an average of 24.707 while our underpredictions were off by an average of -106.469. We see that our worst predictions tend to be in August of each year. Because our test data set is relatively small, it is hard to determine if this is caused by the first storm of a given hurricane season, or if the problem tends to be worse the year after a relatively light hurricane season. The latter possibility would be an interesting feature to include in future improvements.

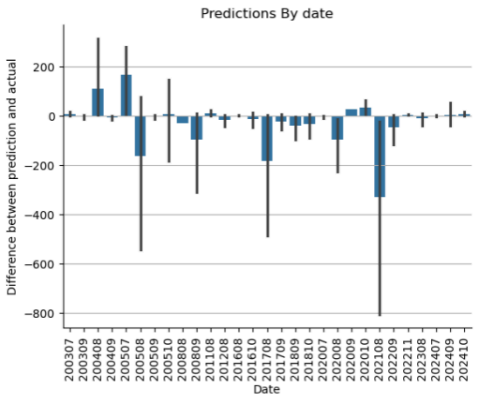


Figure 9 - Predictions by Date

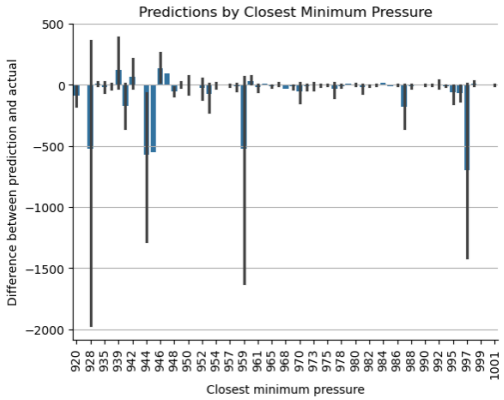


Figure 10 - Predictions by Closest Minimum Pressure

We picked three records for failure analysis – our worst underprediction (Figure 11), overprediction (Figure 12), and our worst overprediction on a target value of zero (Figure 13). SHAP force plots for these errors were generated.

Index 1303 - Target value: 5617, Prediction Error: -3989 units

Figure 11 - Worst underprediction

Index 1306 - Target value: 265, Prediction Error: 716 units

Figure 12 - Worst overprediction

Index 140 - Target value: 0, Prediction Error: -27 units

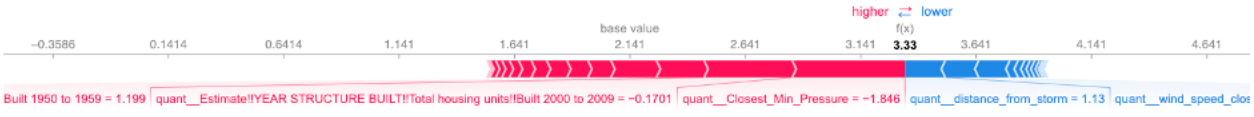


Figure 13 - Worst overprediction on target value of 0

Viewing the first two records, the big variations between the two are in the SHAP values for **wind_speed_closest_div_distance_closest** and **closest_wind_x_pressure_div_distance**, which indicates that our model is unable to assign the correct values to those features to provide accurate predictions. It also could be that the value for **Closest_Min_Pressure,** which is equal for the two predictions, does not provide a proper baseline value upon which our other features can build to make an accurate prediction. As shown in FIgure 10, there are drastic underpredictions of the target at multiple values of **Closest_Min_Pressure**, meaning that this feature is not accurately weighted for successful prediction.

Part B – Unsupervised Learning

After selecting the required data from the Census Bureau tables, we combined the data with the FEMA declarations dataset. This ensured that we only included counties that were impacted by storms. Then, we filtered for the top ten states impacted by hurricanes and isolated our analysis to only the counties in those states that were impacted by hurricanes.

After filtering the data, we applied Principal Component Analysis to reduce the dimensions of the dataset. Using a scree plot (FIgures 14 and 15), it was determined to use 4 principal components, which accounted for approximately 85-90% of the variance in each dataset. For more interpretable results, sparse PCA was also used with 4 principal components. Sparse PCA loadings are shown in Figures 16 and 17.

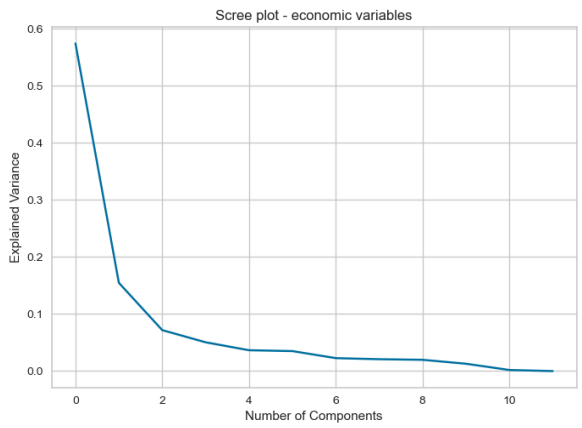


Figure 14 - Scree Plot for Economic Variables

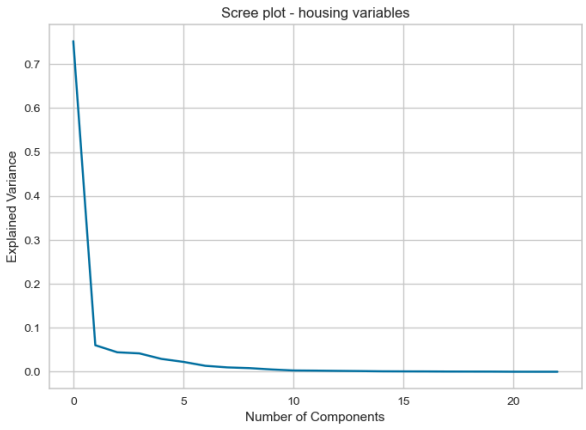


Figure 15 - Scree Plot for Housing Variables

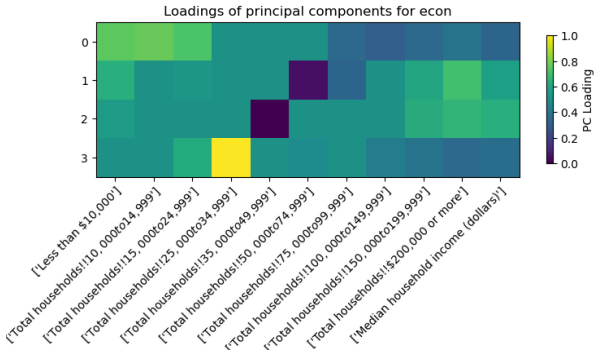


Figure 16 - Sparse PCA Loadings for Economic Variables

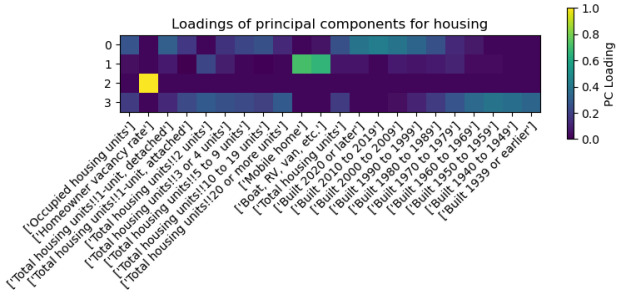


Figure 17 - Sparse PCA Loadings for Housing Variables

For the clustering algorithm, we selected agglomerative clustering and kmeans clustering. We chose not to use DBSCAN or HDBSCAN as we wanted every county assigned to a cluster. Although our model is not going to be deployed, it would not be appropriate to assign residents to a 'noise' county. Additionally, we wanted to include a feature in our supervised learning model for the county cluster. We did not employ spectral clustering since we used PCA, which is linearly separable. Spectral clustering would be more appropriate if clusters were nested, and if we employed kernel PCA with the radial basis function. The kmeans algorithm with parameters of 'k-means++', 10 iterations, and a random state of 42 was used. For agglomerative clustering, we employed the default settings.

A function was created for each clustering method. To compare models, we used Davies-Bouldin Index, Calinski-Harabasz Index, and Silhouette Score as these do not require ground truth labels. The function iterated from 2 to 10 clusters and saved the best scores. The process was repeated for Sparse PCA. For Davies-Bouldin

Index, a measure of cluster separation, a lower score closer to 0 is considered better. A higher score for the Calinski-Harabasz Index, which measures cluster definition with a sum of between cluster dispersion and within-cluster dispersion, is considered better. Silhouette Score, which measures definition using both the mean distance between a sample and its other class members as well as a sample and all points in the next nearest cluster, is bounded between -1 and 1. Scores closer to 1 are for highly dense clusters, whereas a score of 0 indicates overlapping clusters, and -1 means incorrect clustering. The final choice was made on a 'best 2 of 3' judgment, since no one model outperformed on all 3 scores. The chosen model is highlighted in yellow.

Table 3 - Clustering with Principal Component Analysis (4 components, 2 clusters)

		Davies-Bouldin	Calinski-Harabasz	Silhouette
Agglomerative	Housing	0.642	1042	0.83
	Economic	0.775	743	0.47
K-means	Housing	0.492	1170	0.883
	Economic	0.878	770	0.385

Table 4 - Clustering with Sparse Principal Component Analysis (4 components, 2 clusters)

		Davies-Bouldin	Calinski-Harabasz	Silhouette
Agglomerative	Housing	0.397	874	0.91
	Economic	0.61	584	0.488
K-means	Housing	0.543	1020	0.874
	Economic	0.87	662	0.388

To visualize the clusters, we plotted them in terms of their first two principal components, and as a practical matter on choropleth maps of hurricane affected counties.

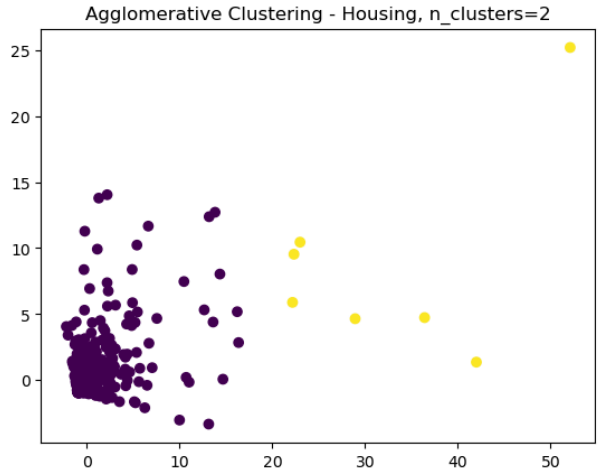


Figure 18 - Agglomerative Housing Clusters plotted by first two PCs

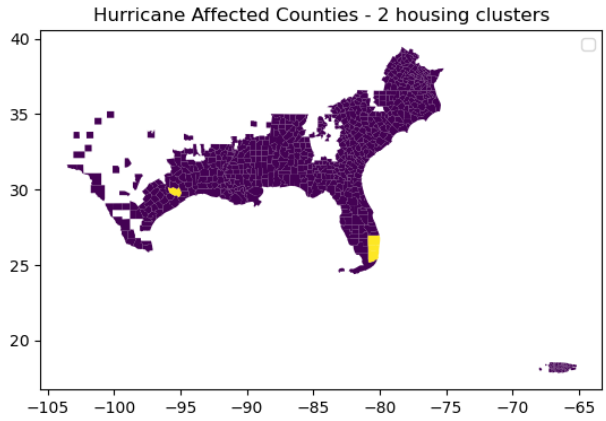


Figure 19 - Housing clusters plotted on map of affected counties

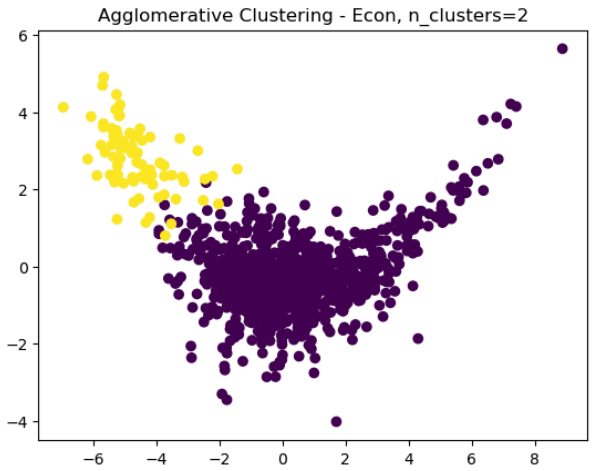


Figure 20 - Agglomerative Econ Clusters plotted by first two PCs

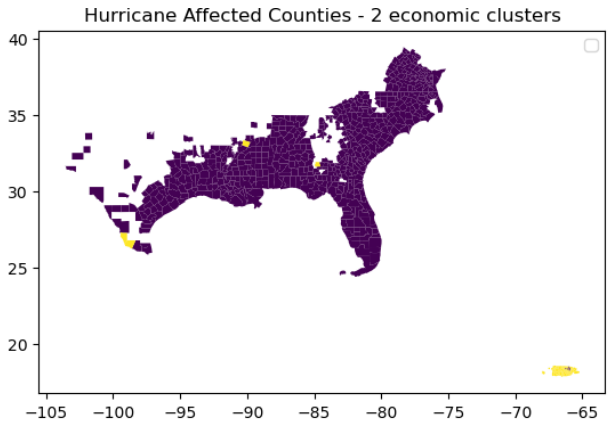


Figure 21 - Econ clusters plotted on map of affected counties

The choropleth maps demonstrate that the clustering was inappropriate. Economic clustering seems to highlight a border county in Texas, Puerto Rico, and another county in Mississippi. Housing seems to separate Houston and Miami, while placing all other counties in the same category. This lack of variation may help explain why the clusters did not add predictive power to the supervised learning model. Future clustering could be performed at Block Group level, or by employing more clusters even though the evaluation metrics determined that two clusters were optimal. Comparison with 5 clusters is provided in Appendix B.

Discussion

The most surprising part of the results from Part A was that our model ended up being so conservative in its predictions of storm damage. An initial concern was that the model would err on the side of overestimating damage. It could be that our model did not learn the number of total housing units in a meaningful enough way, but the mean overestimation was only 24.707 while the mean underestimation was -106.469. We were also surprised at the lack of hyperparameter sensitivity in the models that performed best when tested at their baselines. There is some truth to the adage that the key to success is in selecting the right model, not necessarily in tuning it.

One large challenge that we did encounter was in the skewness of our target. Running the initial R2 tests on the untransformed target created some scary results. Nearly every score resulted in a negative number. Kingsley quickly decided we should perform a log transformation on the target and that provided us some relief when the R2 scores turned positive. The other major challenges occurred when trying to work with only the details from the storm taken at landfall. Again, all tests on this initial dataset were very negative. Scott engineered a way to find information about each storm and its closest pass to a given county and created new features from that, and this provided much better results.

With additional time and resources, we could get to a more granular level with the information and go beyond the county level to the city or block level. Additionally, we could spend more time on a deep learning model that may be able to find hidden connections between our features and the target that we were unable to determine on our own.

The results from Part B were surprising in how the clusters were arranged when plotted. There was a large disparity with the housing clusters, with many of the counties far away from other points. DBSCAN may have been a better fit, but as noted earlier we wanted a hard partition where no county was considered as ‘noise’. One of the challenges we encountered was that not all of the counties were included in the economic data. This is

surprising as 5-year estimates should include all counties. When we exported the results to a CSV file, we had to employ a 'left' join on the housing dataset to ensure that all housing clusters were retained, at the expense of having empty values for 25 counties regarding their economic cluster. One possible extension of our solution would be to create clusters for each year, similar to how we calculated the number of structures for our supervised learning features. This could account for how counties changed over time. Another possibility, with more time and much more resources, would be to possibly cluster based on building codes or materials used. We could even cluster by geographical features such as coastline, soil type, impervious surface or other features that could account for damage other than from wind. However, we did not have easily accessible datasets for these features, so many would have to be created through original research

What surprised us about evaluating Part B was that the evaluation metrics determined that 2 clusters were appropriate for both categories. Initially, we conducted exploration with 5 clusters (see additional maps in Appendix B). This was done with the assumption that economic levels would mean more stratification by economic group, and due to the large disparity in number of structures in urban versus rural areas. However, rigorous use of evaluation metrics determined that two clusters were optimal. One of the challenges was finding PCA and clustering that produced sensible results. One exploration was made with kernel PCA on the assumption that the data was not linearly separable. Spectral clustering was then applied to both the housing and economic datasets. The clusters for economic data were located in only three locations on a scatter plot, which made interpretation challenging. Additionally, the first principal component accounted for greater than 99% of variance in both datasets. This made a 2D plot difficult to accomplish. Due to interpretability, this approach was not pursued further.

8 - Ethical Considerations

Our ethical examination of our project was guided by the Data Ethics Canvas (https://theodi.org/documents/469/Data-Ethics-Canvas-English-Grey-1.pdf) developed by the Open Data Institute. Responses from our review of the applicable sections of the canvas are as follows:

- Data sources: All of our data is publicly available and no personally identifiable or sensitive data is included. Thus, no privacy issues are involved. Further, no protected classes are explicitly included in our dataset.
- Rights around data sources: All of the data used are publicly available via US government sources.
- Limitations in data sources: Some housing data were not available for all years of interest. Thus, interpolation and forecasting techniques were used to produce these missing values. This could produce some inaccuracies and bias predictions. *Also, without proper retraining over time, the model will fail to properly reflect the current regional characteristics and provide increasingly poor results.*
- Ethical and legislative context: There are no unique ethical, regulatory, or legislative requirements applicable to our specific project context (i.e., predicting residential housing damage in the aftermath of a hurricane).
- Your reason for using data: Our intent in using the data is to enable better prediction of hurricane storm damage to residential housing, which has the potential to increase governmental preparedness and responsiveness in the face of hurricanes through better allocation of resources. This provides a benefit to society.
- Positive effects on people: If successful, our model could prove useful for FEMA and storm-affected citizens and speed up recovery efforts in the wake of a hurricane event. Given the data used, the model primarily applies to storms originating in the Atlantic Ocean. Information gleaned from the study might provide evidence in support of modern residential building standards.

- Negative effects on people: The resulting model could be used for unintended purposes such as the establishment of homeowners insurance rates, leading to future provenance issues. Further, given that the intent of our model is to predict *occupied housing damage* at the county level, a clear distinction must be made between this outcome measure and predicting the *number of individuals* impacted at the county level. Thus, there is a risk that the model could be improperly used to allocate non-fiduciary resources such as food, water, and toiletries. Also, there is the potential that use of model could result in an overemphasis on aiding areas likely to be hardest hit in a manner that might detract from efforts needed in other areas suffering damages but in lower quantities.
- Minimizing negative impact: To help mitigate the potential negative effects on people, the specific intent of the model must be communicated clearly as well as the limitations associated with attempts to use the model outside of its intended context, including outside of the applicable geographic regions. With the understanding that the county data used in our model represent real citizens of the US, we were careful to avoid the use of unsupervised learning models that would categorize any counties as mere noise as we deemed this inappropriate.
- Engaging with people: Given the academic nature of this project, this was deemed not applicable.
- Communicating your purpose: Given the academic nature of this project, communication is limited to this report.
- Openness and transparency: Given the academic nature of this project, this was deemed not applicable.
- Sharing data with others: Given the academic nature of this project, this was deemed not applicable. There is no plan to share this model outside of the participants in this course.
- Ongoing implementation: Given the academic nature of this project, this was deemed not applicable.
- Reviews and iterations: Given the academic nature of this project, this was deemed not applicable beyond the reviews and iterations internal to the team to develop the best model based on selected evaluation metrics.
- Your actions: Given the academic nature of this project, this was deemed not applicable.

Given the above, we feel that we have a solid ethical foundation regarding our project aims and execution.

Statement of Work

Scott Powell	Kingsley A. Reeves, Jr.	Joe Schweiss
Project Management, data sources identification, FEMA data extraction and cleaning, final dataset merging and cleaning, unsupervised learning modeling and cluster analysis, ethical analysis, report writing	Hurricane data extraction and cleaning, census data extraction and cleaning, supervised learning modeling, ethical analysis, report writing	Supervised learning modeling, failure analysis, sensitivity analysis, error analysis, ethical analysis, report writing, GitHub handling

References

AOML Communications. “New NOAA Research Predicts an Increase in Active Atlantic Hurricane Seasons.” NOAA’s Atlantic Oceanographic and Meteorological Laboratory, 18 Nov. 2024, www.aoml.noaa.gov/new-noaa-research-predicts-an-increase-in-active-atlantic-hurricane-seasons/.

Hosmay Lopez et al., “Projected Increase in the Frequency of Extremely Active Atlantic Hurricane Seasons.” *Science Advances* 10.46, (2024).DOI:10.1126/sciadv.adq7856

Klepac, S et al. “A Case Study and Parametric Analysis of Predicting Hurricane-Induced Building Damage Using Data-Driven Machine Learning Approach.” *Frontiers in Built Environment.* 8 (2022): n. pag. Web.

Li, Y & Gu, S. "Detecting Post Hurricane House Damage Using Geographic Information Related Multi-Resource Classification Model," 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Zhuhai, China, 2021, pp. 492-501, doi: 10.1109/ICBASE53849.2021.00098.

Ma, CY et al. “Vulnerability of Renters and Low-Income Households to Storm Damage: Evidence From Hurricane Maria in Puerto Rico.” *American Journal of Public Health.* 110.2 (2020): 196–202. Web.

Pilkington, SF & Mahmoud, HN. “Using Artificial Neural Networks to Forecast Economic Impact of Multi-hazard Hurricane-based Events.” *Sustainable and Resilient Infrastructure*, *1*(1–2), (2016)63–83. https://doi.org/10.1080/23789689.2016.1179529

Pinelli, JP et al. “Hurricane Damage Prediction Model for Residential Structures.” *Journal of Structural Engineering.* 130.11 (2004): 1685–1691. Web.

Appendix A - Features for Supervised Learning

Damage	Total number of damaged houses at the county level (**target variable**)	Integer	Engineered from Federal Emergency Management Agency (FEMA) Housing Assistance Program Data
lnDamage	The natural log of the number of damaged houses was computed due to the skewed distribution of the untransformed data (**transformed target variable**)	Float	Engineered from Federal Emergency Management Agency (FEMA) Housing Assistance Program Data
Estimate!!HOUSING OCCUPANCY!!Total housing units	Total housing units at the state and county levels	Integer	Engineered from US Census Bureau

			American Community Survey (ACS) DP04 Selected Housing Characteristics 2010, 2015, 2019, and 2023 5-Year Estimates and 2000 Decennial Housing Data
Estimate!!HOUSING OCCUPANCY!!Total housing units!!Occupied housing units	Total occupied housing units at the state and county levels	Integer	Engineered from US Census Bureau American Community Survey (ACS) DP04 Selected Housing Characteristics 2010, 2015, 2019, and 2023 5-Year Estimates and 2000 Decennial
Estimate!!YEAR STRUCTURE BUILT!!Total housing units!!Built 2020 or later	Total housing units built in 2020 or later at the state and county levels	Integer	Engineered from US Census Bureau American Community Survey (ACS) DP04 Selected Housing Characteristics 2010, 2015, 2019, and 2023 5-Year Estimates and 2000 Decennial data
STRUCTURE BUILT!!Total housing units!!Built XXXX to YYYY	Total housing units built between XXXX and YYYY at the state and county levels	Integer	Engineered from US Census Bureau American Community Survey (ACS) DP04 Selected Housing Characteristics 2010, 2015, 2019, and 2023 5-Year Estimates and 2000 Decennial data
Estimate!!YEAR STRUCTURE BUILT!!Total housing units!!Built 1939 or earlier	Total housing units built in 1939 or later at the state and county levels	Integer	Engineered from US Census Bureau American Community Survey (ACS) DP04

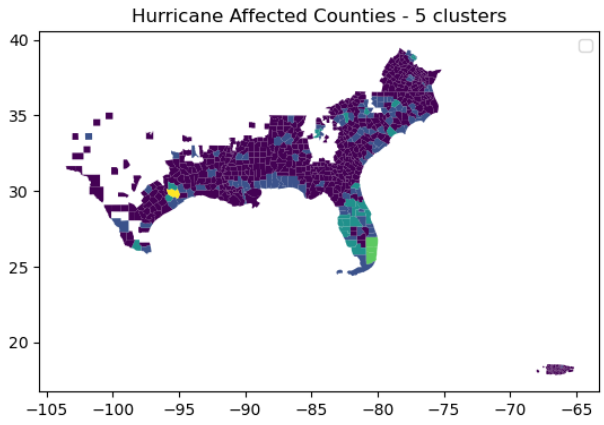
			Selected Housing Characteristics 2010, 2015, 2019, and 2023 5-Year Estimates and 2000 Decennial data
Max Wind	Maximum sustained wind speed (in knots) achieved by a storm at landfall in region of interest	Integer	National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
Min Pressure	Minimum pressure (in millibars) achieved by a storm at landfall in region of interest	Integer	National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
dist_from_landfall	Regional center's distance from the storm (in miles) at the storm's landfall location (in miles, as a proxy for storm strength in the affected region)	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
Closest_Max_Wind	Maximum sustained wind speed (in knots) achieved by a storm at its closest pass to the affected region	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic

			hurricane database data (HURDAT2)
Closest_Min_Pressure	Minimum pressure (in millibars) achieved by a storm at its closest pass to the affected region	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
distance_from_storm	Regional center's distance from the storm (in miles) at the storm's closest pass (as a more refined proxy for storm strength in the affected region)	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
category	Categorizes as a tropical depression, tropical storm, or level 1-5 hurricane	Categorical	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
housing_cluster	A categorical identifier grouping similar counties together base on housing characteristics	Categorical	US Census Bureau American Community Survey (ACS) DP04 Selected Housing Characteristics 5-Year Estimates data

econ_cluster	A categorical identifier grouping similar counties together based on economic characteristics	Categorical	US Census Bureau American Community Survey (ACS) CP03 Selected Economic Characteristics 5-Year Estimates data
max_wind_speed_div_distance	A second order measure of storm strength resulting from the quotient of Max Wind and dist_from_landfall	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
max_wind_speed_div_distance_pressure	A third order measure of storm strength resulting from Max Wind divided by the product of dist_from_landfall and Min Pressure	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
wind_speed_closest_div_distance_closest	A second order measure of storm strength resulting from the quotient of Closest Max Wind and distance_from_storm	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
max_wind_x_pressure	A second order measure of storm strength resulting	Float	Engineered using Census Bureau location data and

	from the product of Max Wind and Min Pressure		National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
max_wind_x_pressure_div_distance	A third order measure of storm strength resulting from the product of Max Wind and Min Pressure, divided by dist_from_landfall	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)
closest_wind_x_pressure_div_distance	A third order measure of storm strength resulting from the product of Closest_Max_Wind and Closest_Min_Pressure, divided by distance_from_storm	Float	Engineered using Census Bureau location data and National Oceanic and Atmospheric Association (NOAA) National Hurricane Center (NHC) Atlantic hurricane database data (HURDAT2)

Appendix B - Supplemental Maps



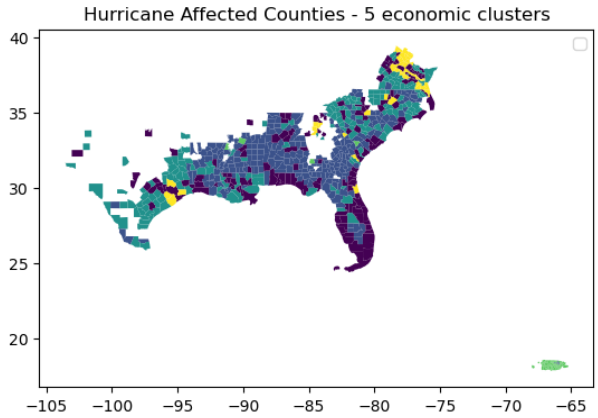


Figure 22 - Housing clusters with 5 clusters

Figure 23 - Econ clusters with 5 clusters

The five arbitrarily chosen clusters provided some practical, and expected, separation. However, this was not deemed optimal based on our employment of clustering evaluation metrics. When implemented in our model, the 5 clusters did not improve performance, indicating that they were likely inappropriate to begin with.