When the Levees Broke: Adding Socioeconomic Dimensionality to Flood Risk Predictive Modeling

Sathya Edamadaka, Marie-Claire Traore, Will Walker, Kristen Ray

Highlights

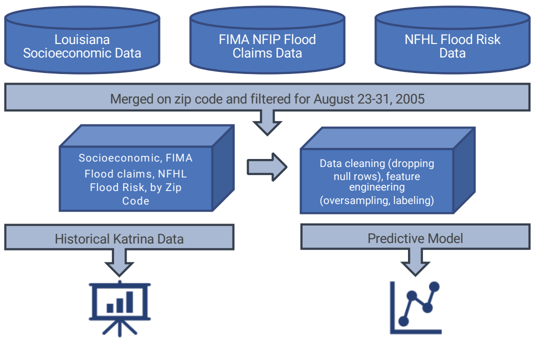
- Our approach uniquely considers socioeconomic data alongside geographical risk to develop a better picture for flood risk prediction during large storms. We considered Louisiana during Hurricane Katrina.
- Our artificial neural network and XGBoost-based models achieve accuracies of 86%, as well as identifying chronically underinvested areas like the Lower Ninth Ward.
- Our predictive models can better support the management and design of coastal infrastructure to minimize damage caused by tropical storms in Louisiana.

Context

Since 1980, tropical storms have accounted for $945.9 billion in damages. Due to **climate change**, the frequency of **stronger**, more **destructive** tropical storms has continued to increase . Additionally, climate change has caused an increase in storms that **rapidly intensify** just before landfall, making it difficult for communities to adequately prepare for an approaching storm.

Data

By looking at how Hurricane Katrina impacted Louisiana, this model examines the factors by zip code and considers local socioeconomic data, FIMA flood claims data and NFHL flood risk data.



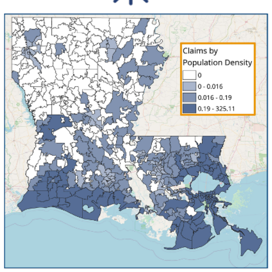


Figure 1. NFIP Flood count data, normalized by population density

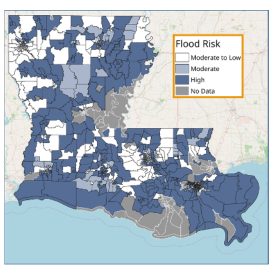


Figure 2. NFHL Flood risk factors

Preliminary Modeling

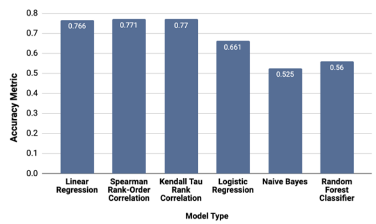


Figure 3. We initially applied a variety of simpler modeling techniques to our dataset—a number of linear regression methods achieved high accuracy, while more complicated naive bayesian and random forest classifiers performed poorly.

Oversampling Method

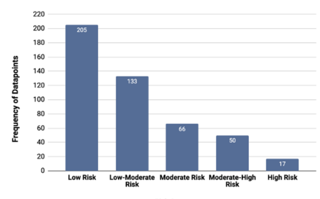


Figure 4. Original risk category distribution— very little data for all non-low risk categories

Borderline SMOTE

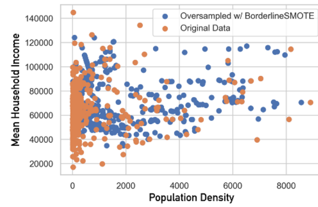


Figure 5. Final, oversampled data. Mean household income & population density shown as examples

Instead of just analyzing geographical flood risk factors, we also analyze areas in Louisiana through socioeconomic and demographic statistics. By doing so, we predicted which areas will be heavily impacted in future storms, identifying areas with disproportionately poor infrastructure along the way (e.g. the Lower Ninth Ward).

Artificial Neural Network and XGBoost Modeling

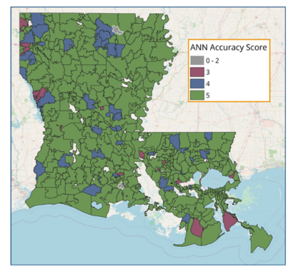
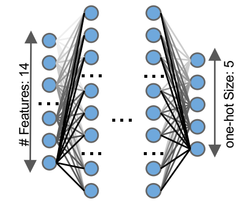


Figure 6. Accuracy plot for ANN—an accuracy score of 5 is a perfect guess, and 0 is completely wrong, as shown in the accuracy score formula:

$$\text{score} = 5 - \left|\text{risk}_{\text{predicted}} - \text{risk}_{\text{actual}}\right|$$

84% Accuracy, F1-score, Recall, and Precision



If you're interested, here are some model specifics: weights were updated with Adam optimization. Rectifying linear unit was used for our activation function. We used a constant learning rate of 10^{-5}. We found two hidden layer structures that were highly accurate and that didn't overfit: (41, 31,41, 21, 51) and (91, 71, 81).

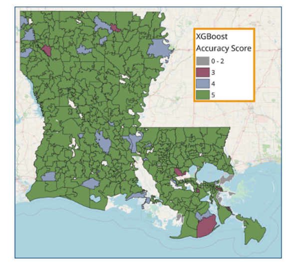


Figure 7. Accuracy plot for XGBoost, using the same metric as before. It's interesting to note that this model is wrong in different places than the ANN, and is either absolutely correct or very wrong in the majority of LA zip codes.

87% Accuracy, F1-score, Recall, and Precision

Here's a truncated version of the XGBoost decision tree. It's based off of making a series of decisions, informed by previous iterations of the algorithm. It's more accurate than ANN but the magnitude of its errors are larger.

Dashboard Highlights

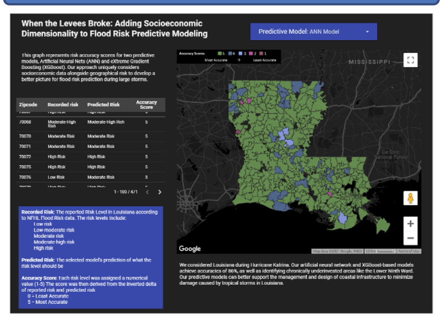
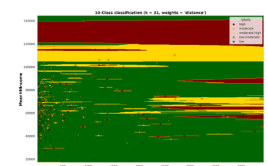


Figure 8. Data dashboard representing eXtreme Gradient Boosting (XGBoost) and Artificial Neural Network (ANN) predictive models, developed with Google Data Studio

K-Nearest Neighbors Modeling



78% Accuracy, F1-score, Recall, and Precision

Figure 9. This is a 2D plot of the most highly weighted variables, population density and mean household income, after training the 14-dimensional kNN model. The important thing to take away here is that this diagram is not simple— as a result, we had to utilize the above, nonlinear methods for analysis.

Conclusion

By taking socioeconomic and demographic factors into account, our models highlight that factors other than geographical features can be used to determine the likelihood of flooding in a region hit by a tropical storm. We hope our results can be used to inform infrastructure planning in cities to reduce the impact of natural disasters, guiding future investment and planning.

When the Levees Broke: Adding Socioeconomic Dimensionality to Flood Risk Predictive Modeling

Sathya Edamadaka, Marie-Claire Traore, Will Walker, Kristen Ray

Highlights

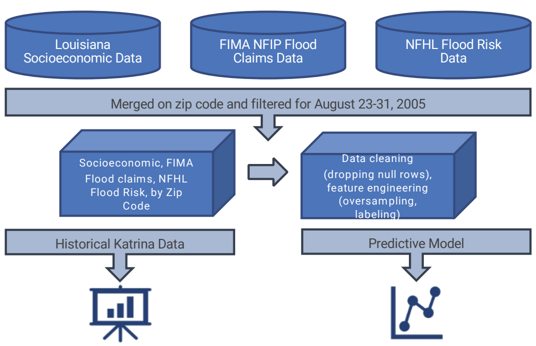
- Our approach uniquely considers socioeconomic data alongside geographical risk to develop a better picture for flood risk prediction during large storms. We considered Louisiana during Hurricane Katrina.
- Our artificial neural network and XGBoost-based models achieve accuracies of 86%, as well as identifying chronically underinvested areas like the Lower Ninth Ward.
- Our predictive models can better support the management and design of coastal infrastructure to minimize damage caused by tropical storms in Louisiana.

Context

Since 1980, tropical storms have accounted for $945.9 billion in damages. Due to **climate change**, the frequency of **stronger**, more **destructive** tropical storms has continued to increase . Additionally, climate change has caused an increase in storms that **rapidly intensify** just before landfall, making it difficult for communities to adequately prepare for an approaching storm.

Data

By looking at how Hurricane Katrina impacted Louisiana, this model examines the factors by zip code and considers local socioeconomic data, FIMA flood claims data and NFHL flood risk data.



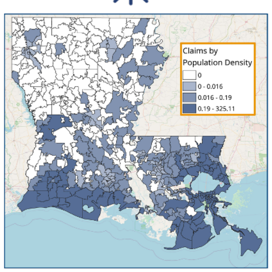


Figure 1. NFIP Flood count data, normalized by population density

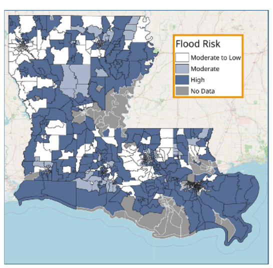


Figure 2. NFHL Flood risk factors

Preliminary Modeling

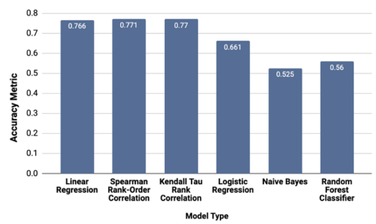
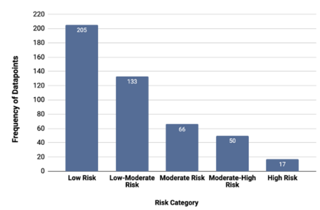
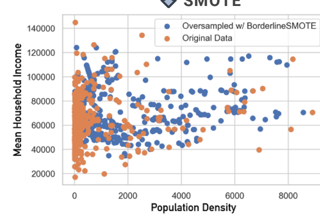


Figure 3. We initially applied a variety of simpler modeling techniques to our dataset—a number of linear regression methods achieved high accuracy, while more complicated naive bayesian and random forest classifiers performed poorly.

Oversampling Method

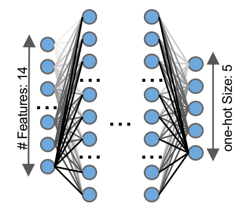


Borderline SMOTE

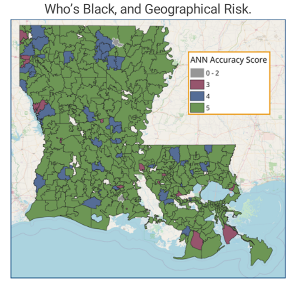


Instead of just analyzing geographical flood risk factors, we also analyze areas in Louisiana through socioeconomic and demographic statistics. By doing so, we predicted which areas will be heavily impacted in future storms, identifying areas with disproportionately poor infrastructure along the way (e.g. the Lower Ninth Ward).

Artificial Neural Network and XGBoost Modeling

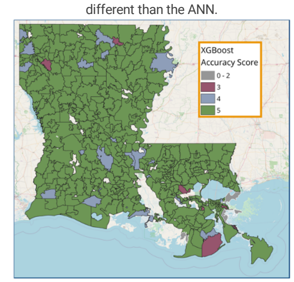


Artificial Neural Network:
84% Accuracy.
This model mostly used Density, Median Household Income, Proportion of Population Who's Black, and Geographical Risk.



Accuracy plot for ANN—an accuracy score of 5 is a perfect guess, and 0 is completely wrong. Please notice that the Lower 9th Ward was guessed correctly.

XGBoost-Based Model:
87% Accuracy.
The model mostly used Density, Total Population, and Bachelor's Rate, much different than the ANN.



Accuracy plot for XGBoost, using the same metric as before. It's interesting to note that this model is wrong in different places than the ANN, and is either absolutely correct or very wrong in the majority of LA zip codes.