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

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

Our Team

Will Walker

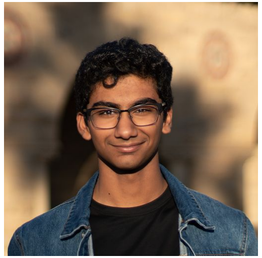
Data Analyst Intern at Technology Rediscovery, Temple University 2023 (Data Science/ Cognitive Neuroscience)

Marie-Claire Traore

CODA Associate at Capital One

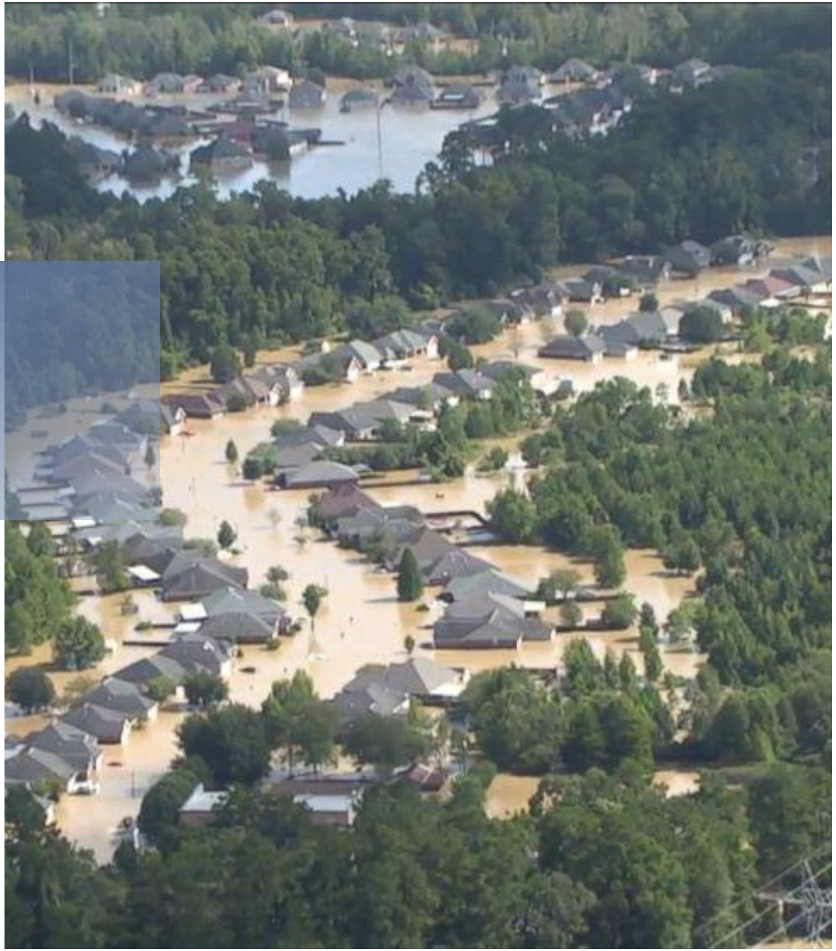
Kristen Ray

Cognitive Data Consultant for IBM



Sathya Edamadaka

SWE at BuildOps and Researcher at SLAC, Stanford 2023 (EE & Physics)



Problem Statement

- Since 1980, tropical storms have accounted for $945.9 billion in damages surpassing all other weather-related disasters in the US
- Due to climate change, the destruction caused by tropical storms continues to increase
- We developed a predictive model to help support the management and design of coastal infrastructure to minimize damage caused by tropical storms.

Our Approach

- We decided to add socioeconomic and demographic data to traditional geographic flood risk factors.

- By doing so, we are able to predict the level of impact each area had during Hurricane Katrina, as well as identify areas with disproportionately poor infrastructure

Modeling Timeline

Gathering Data

Socioeconomic data (Census)
Geographical risk data (NFHL)
Flood damage report data (NFIP)

Cleaning Data

471/474 Zip Codes

Feature Selection

Feature Engineering

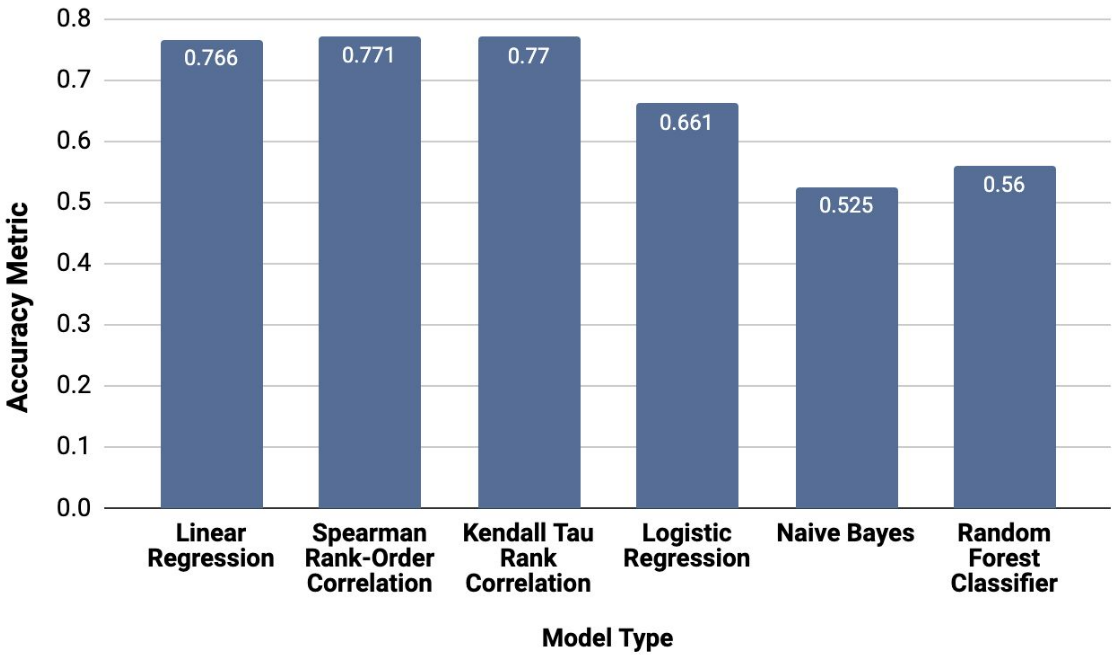
Each zip code labeled by how many residents per square mile reported damages:

- Low, low-moderate, ... , high risk

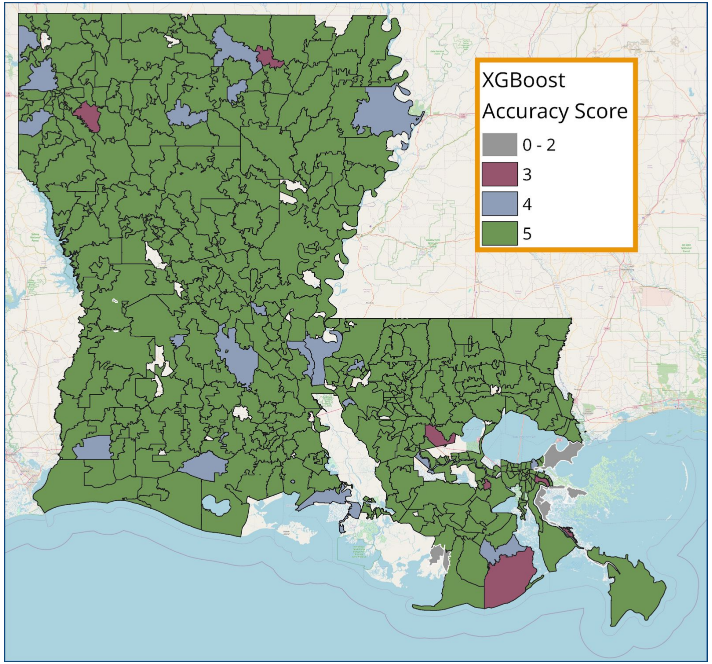
Building and Testing Models

Tried both simple and complex modeling

Initial Model Performance — Good, but not great



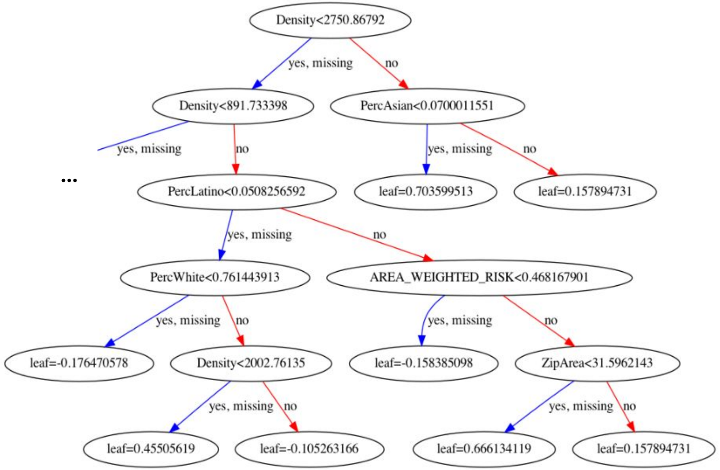
XGBoost Modeling



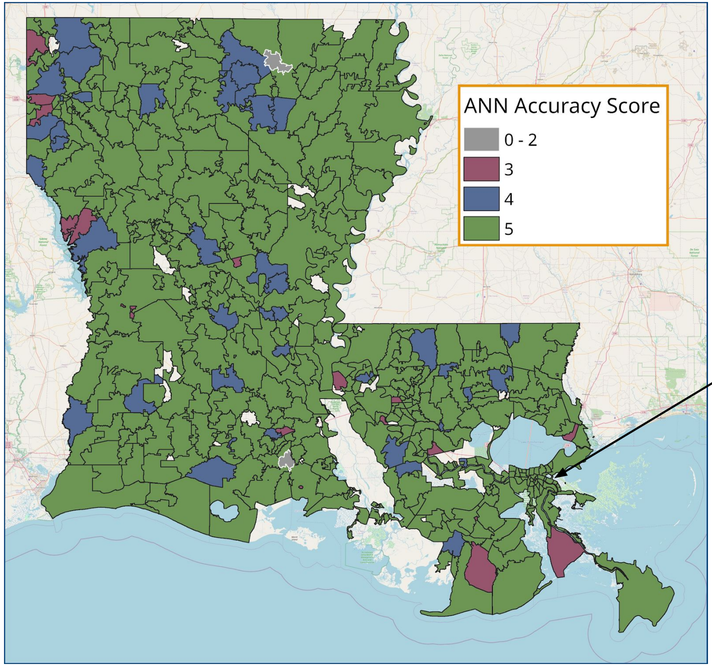
$$\text{score} = 5 - \left|\text{risk}_{\text{predicted}} - \text{risk}_{\text{actual}}\right|$$

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

Most important features: Density, Total Population, and Bachelor’s Rate



Artificial Neural Network (ANN) Modeling

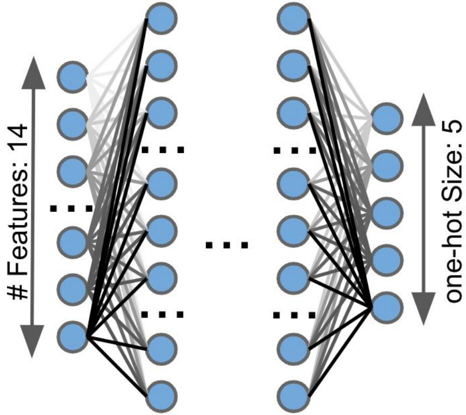


Lower 9th Ward correctly predicted!

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

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

Most Important Features: Density, Median Household Income, Proportion of Population Who's Black, Geographical Risk



Data Dashboard

Link to Dashboard:

https://datastudio.google.com/reporting/d1b74776-4a0e-48b8-a5c8-fc41016b32a2

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

Predictive Model: ANN Model

This graph represents risk accuracy scores for two predictive models, Artificial Neural Nets (ANN) and eXtreme Gradient Boosting (XGBoost). Our approach uniquely considers socioeconomic data alongside geographical risk to develop a better picture for flood risk prediction during large storms.

Zipcode	Recorded risk	Predicted Risk	Accuracy Score
70068	Moderate-High Rish	Moderate-High Rish	5
70070	Moderate Risk	Moderate Risk	5
70071	Moderate Risk	Moderate Risk	5
70072	High Risk	High Risk	5
70075	High Risk	High Risk	5
70076	Low Risk	Moderate Risk	3

1 - 100 / 471

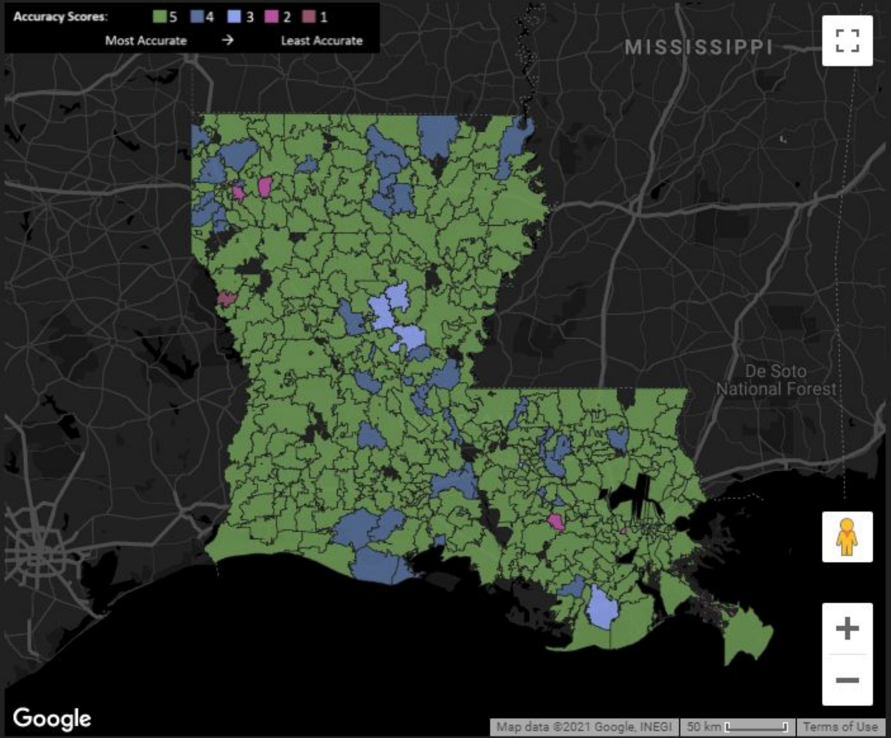
Recorded Risk: The reported Risk Level in Louisiana according to NFHL Flood Risk data. The risk levels include:

- Low risk
- Low-moderate risk
- Moderate risk
- Moderate-high risk
- High risk

Predicted Risk: The selected model's prediction of what the risk level should be

Accuracy Score: Each risk level was assigned a numerical value (1-5) The score was then derived from the inverted delta of reported risk and predicted risk

- 0 = Least Accurate
- 5 = Most Accurate

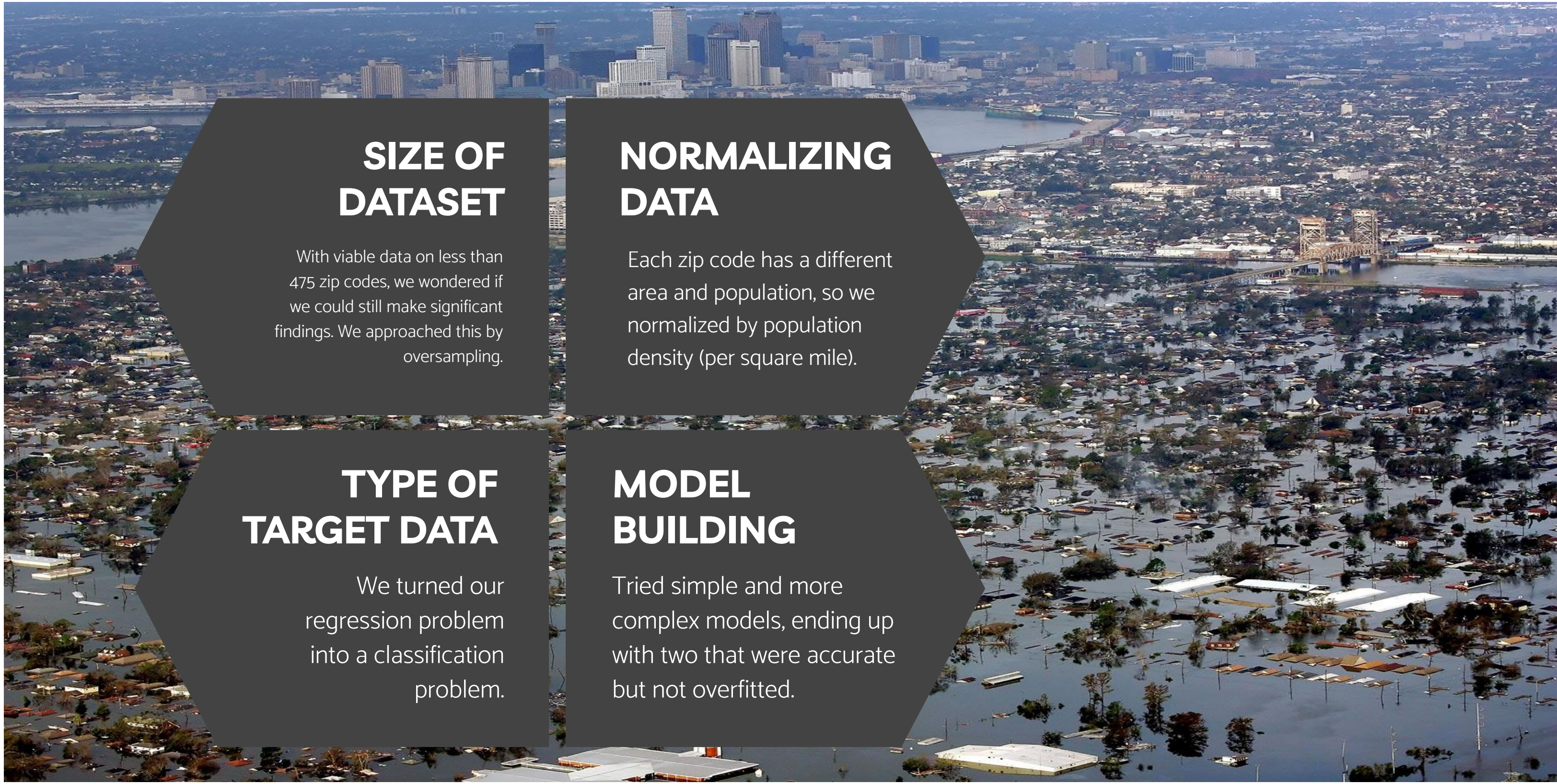


We considered Louisiana during Hurricane Katrina. Our artificial neural network and XGBoost-based models achieve accuracies of 86%, as well as identifying chronically underinvested areas like the Lower Ninth Ward. Our predictive models can better support the management and design of coastal infrastructure to minimize damage caused by tropical storms in Louisiana.

Conclusion

- Significant impact on the infrastructure considerations of Louisiana
- Improving quality of life and safety of those in flood zones

Recap of Project Challenges



SIZE OF DATASET

With viable data on less than 475 zip codes, we wondered if we could still make significant findings. We approached this by oversampling.

NORMALIZING DATA

Each zip code has a different area and population, so we normalized by population density (per square mile).

TYPE OF TARGET DATA

We turned our regression problem into a classification problem.

MODEL BUILDING

Tried simple and more complex models, ending up with two that were accurate but not overfitted.

Future Work

Expanding to Different Storms and Locations

As this work focused on Louisiana during the time of Hurricane Katrina, we'd love to expand to looking at other Gulf Coast states like Texas, Florida and Mississippi during other tropical storms of varying intensity.

In-Depth Feature Analysis

We'd love to rerun our analysis with nonlinear transformations to our variables–linear regression may get us really far!

Further Modeling

We'd love to try different models to improve analysis–using CNNs could improve our performance while preventing from overfitting!

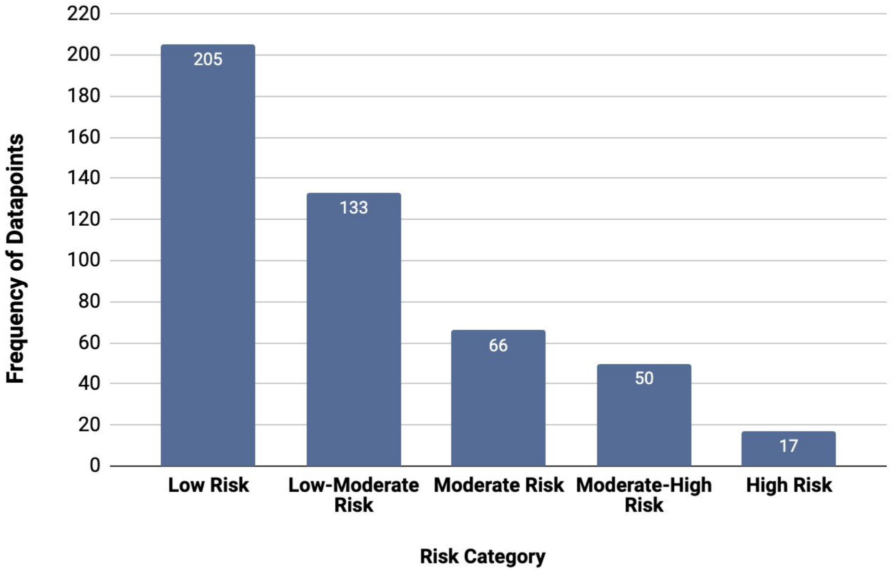
Thank You!

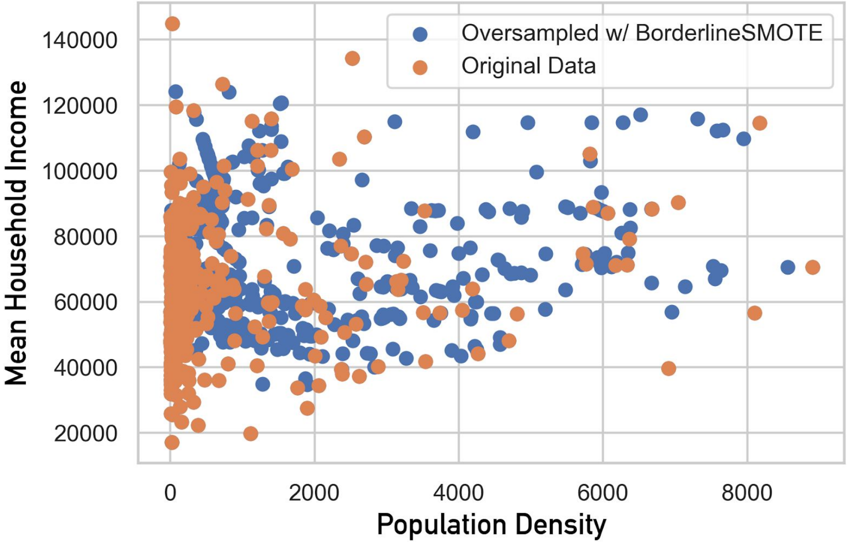
Do you have any questions?

Github: https://github.com/snme/DS4A
Project Dashboard:
https://datastudio.google.com/u/0/reporting/d1b74776-4a0e-48b8-a5c8-fc41016b32a2/page/uR20B

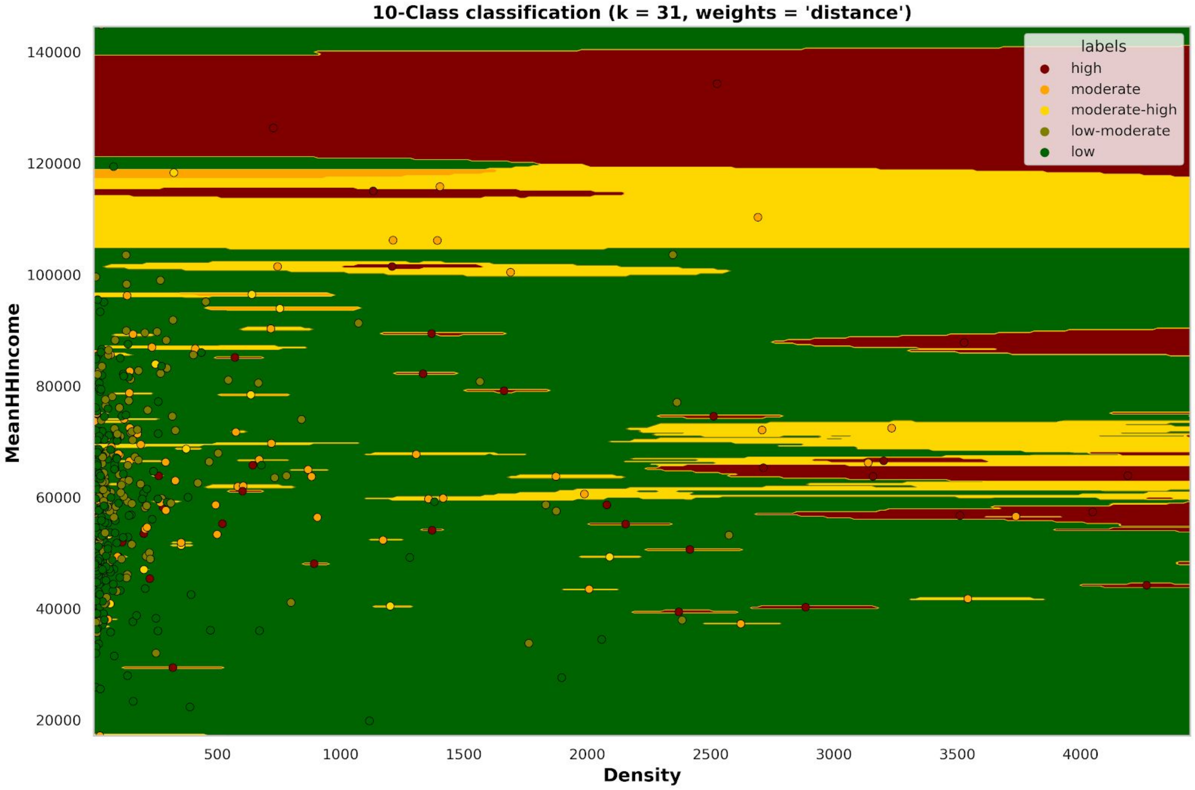
Special thanks to our mentors (Chuck Ni and Raul Aguilar), and our TA Bethini Williams!

Backup - Oversampling (Borderline SMOTE)





Backup - K-Nearest Neighbors



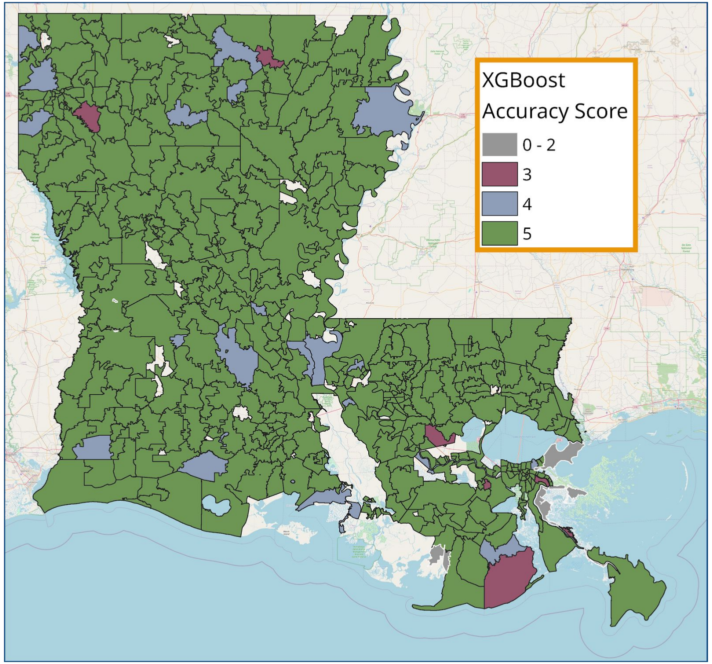
78% Accuracy, F1-score, Recall, and Precision

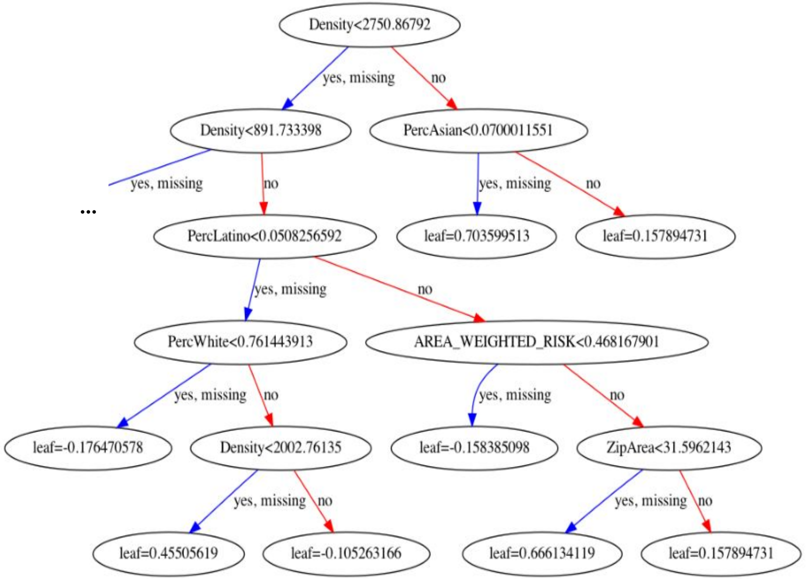
← (it's complicated!)

XGBoost Modeling

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

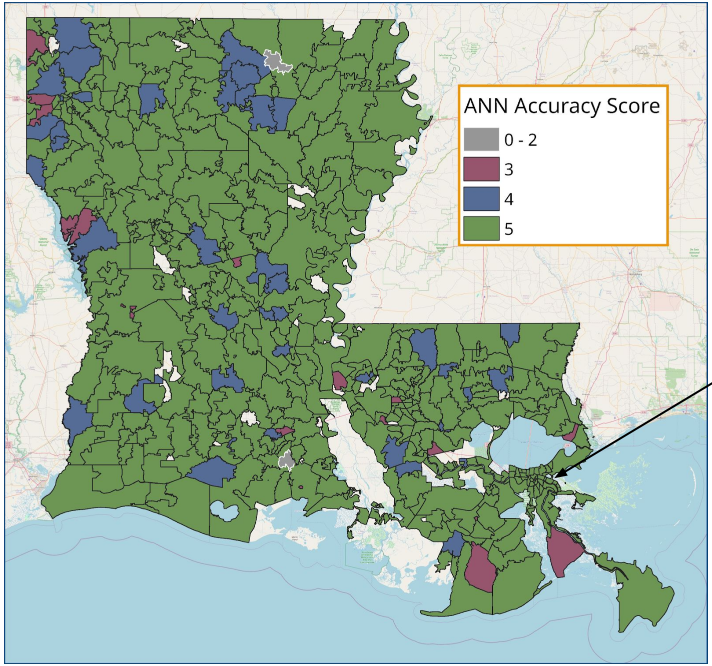
Most important features: Density, Total Population, and Bachelor's Rate





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

Artificial Neural Network (ANN) Modeling



Lower 9th Ward correctly predicted!

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

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

Most Important Features: Density, Median Household Income, Proportion of Population Who's Black, Geographical Risk