From Analysis to Prediction: Examining Socioeconomic Factors and Machine Learning Models in Hurricane Preparedness

Group 10: Tianle Duan, Richard Storm, Prachi Divij Tanna, Ajuna Puthenpurackal John

1 Problem Statement

The challenge is to understand which socioeconomic factors influence a community's preparedness before a hurricane. Many communities are repeatedly affected by hurricanes, yet levels of preparedness vary significantly among different counties. Identifying the key socioeconomic factors that predict preparedness can help allocate resources more effectively and implement targeted interventions.

This project has two main research questions: (1) Which socioeconomic factor(s) have statistically significant contribution to a community's preparedness for an impending hurricane; and (2) Is there any effective machine learning algorithms can accurately predict the decision for hurricane preparedness? To answer the questions, we select county resolution as our research scale. The hypothesis for the first research question is that county in high socioeconomic status and high mobility is associated with proactive preparedness. In other words, counties with higher median income, higher education attainment, more proportion of people in working ages tend to have a statistically significant proactive preparedness pattern.

Understanding the socioeconomic factors that influence preparedness is crucial for disaster management agencies, policymakers, and local governments to develop effective preparedness strategies and policies. Accurate identification of these factors can help prioritize vulnerable communities and allocate resources effectively, potentially saving lives and reducing economic losses.

Solving this problem will provide insights for emergency management professionals to enhance hurricane preparedness programs. It will help target high-risk communities, improve public safety campaigns, and optimize resource distribution, thereby reducing the overall impact of hurricanes on communities.

2 Goals

2.1 Main Goal:

The primary goal of this project is to analyze and predict the preparedness pattern of counties before a hurricane based on demographic factors. More specifically, we will answer the research

questions discussed above: To find the parameters with statistically significant contributions to the preparedness pattern, and find the best machine learning model to predict the preparedness.

The main goal can be divided into some specific goals:

Goal 1: Using logistic regression to understand the impact of various demographic variables (such as median income, percentage of white population, percentage of households without cars, etc.) on the likelihood of a county being prepared for a hurricane. This will help identify which factors are significant predictors of preparedness.

Goal 2: To apply multiple machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machines (GBM), to build predictive models that can accurately classify county based on their preparedness.

Goal 3: To compare the performance of logistic regression with other machine learning models using appropriate metrics (such as accuracy, precision, recall, F1-score, and AUC-ROC) to determine the most effective model for predicting preparedness.

Goal 4: To provide actionable insights for policymakers and emergency management professionals by identifying key socioeconomic factors influencing preparedness and by recommending the best predictive model for identifying high-risk communities.

Goal 5: For us, to better understand machine learning methods and its application.

2.2 Success Metrics:

The success of Goal 1 will be measured by the statistical significance (p-values) and the logistic regression coefficients. The success of Goal 2 and Goal 3 will be evaluated based on model performance metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The best model will be the one with the highest overall performance across these metrics. Goal 4 will be considered successful if the project provides clear, data-driven recommendations for policymakers and emergency management professionals, with evidence of key factors affecting hurricane preparedness and an effective predictive model.

2.3 Relevance to the Problem Statement:

These goals are directly related to the problem statement, which aims to understand the demographic predictors of preparedness and identify the best machine learning model for predicting preparedness. Achieving these goals will provide both interpretative insights and practical solutions for improving hurricane preparedness in vulnerable communities.

3 Data

The data is sourced from two primary datasets: (1) foot traffic data to Points of Interest (POIs), which is processed to determine if a county has a preparedness pattern or not, and (2) socioeconomic data from the U.S. Census Bureau, specifically the American Community Survey (ACS) data which is publicly available.

The response variable is binary, indicating whether a county shows a preparedness pattern (1 for yes, 0 for no). This is determined based on changes in foot traffic data to specific POIs before the hurricane. The explanatory variables include various socioeconomic factors described in Table 1: median income, percentage of white population, percentage of high education attainment, percentage of households without cars, and percentage of households with limited English proficiency.

The dataset includes counties from all counties affected by Hurricane Ida in 2021. Specifically, the analysis will cover **558** counties across multiple states, including Louisiana (LA), Texas (TX), Florida (FL), Mississippi (MS), New York (NY), New Jersey (NJ), and Connecticut (CT). A total of **8** explanatory parameters illustrated in Table 1 are included in the model. The sample size of the data is sufficient enough based on the number of parameters we are going to investigate.

Any missing data in the socioeconomic dataset or foot traffic data will be addressed through appropriate imputation methods or exclusion, depending on the extent and nature of the missing values.

#	Parameters	Unit	Type	Source
Dependent parameters				
1	Preparedness pattern of each COUNTY (yes or no)	\	Binary	Dewey inc.
Independent parameters				
1	Median income	$	Continuous	US census bureau
2	Pct. white population	%	Continuous	US census bureau
3	Pct. people over 65	%	Continuous	US census bureau
4	Pct. people under 5	%	Continuous	US census bureau
5	Pct. people with college degree	%	Continuous	US census bureau
6	Pct. people without high school degree	%	Continuous	US census bureau
7	Pct. households without vehicles	%	Continuous	US census bureau
8	Pct. households with limited English proficiency	%	Continuous	US census bureau

4 Methods:

4.1 Statistical Methods Used

The primary method for understanding the impact of each independent variable on preparedness is Logistic Regression. This method is chosen because it provides interpretable coefficients that reveal the direction and magnitude of the effect of each socioeconomic variable on the binary outcome (preparedness: yes or no).

To find the most accurate predictive model, additional machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machines (GBM), will be applied. These methods are chosen for their ability to handle complex, non-linear relationships and interactions among variables.

4.2 Description of the Methods

Logistic Regression will be used to fit a model where the binary response variable (preparedness or not) is predicted using the selected socioeconomic factors. This method will help identify the most significant predictors and provide insights into how each factor affects the likelihood of county being prepared.

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and aggregates their results to improve prediction accuracy and reduce overfitting. Feature importance scores will be extracted to understand the relative contribution of each variable to the model's prediction.

Support Vector Machine (SVM) will be used with a linear or non-linear kernel (depending on the data structure) to find the optimal hyperplane that separates the prepared and non-prepared COUNTYs, maximizing the margin between the two classes.

Gradient Boosting Machines (GBM), including popular implementations like XGBoost, will be employed to iteratively build an ensemble of weak learners (such as shallow decision trees) to minimize prediction error and boost model performance.

4.3 Methods relevance explanation

Logistic Regression is particularly relevant for its interpretability, which aligns with the first project goal of understanding the impact of socioeconomic variables on preparedness.

Random Forest, SVM, and GBM are selected for their high accuracy and robustness in handling various types of data, addressing the second project goal of finding the most effective predictive model. These methods can capture complex patterns and interactions that logistic regression might miss, providing a comprehensive comparison of model performance across different algorithms.