Michelle Lee
Emily Moberly
Alyssa Hara
Valeria Zheng
Manuel Martinez Garcia

Predictive Modeling for Addressing Homelessness in California

Problem Statement

The issue of homelessness remains critical in the U.S., with many factors contributing to its complexity, such as housing availability, economic shifts, and local policies. Despite numerous efforts, homelessness persists, and the challenge lies in anticipating future trends and areas of high need. While governments, local facilities, and nonprofits dedicate substantial resources to address this issue, homelessness trends can shift due to a wide range of factors, making it difficult to predict where and when resources and funding are most needed. Without accurate predictions, interventions may fall short, leaving certain regions and populations underserved.

Data

The primary dataset for this project is the 2007-2023 Point-in-Time (PIT) Counts by Continuum of Care (CoC). This dataset provides annual estimates of the homeless population across different regions across the United States, including demographic breakdowns and subpopulation counts (e.g. veterans, families, individuals with mental illness). Additionally we also plan on finding data related to:

- Temperature
- Current population density
- Housing cost
- Crime rate
- High school graduation rate
- Average salary
- Unemployment rate
- Political affiliation of county
- Convictions rate
- Substance abuse
- Veteran status
- Poverty rate
- Mental health ratio
- Disability rates
- Adoption, foster system
- Immigrant population
- Funding and welfare
- Count of welfare offices
- Air quality index
- Number of shelters,
- Price of gas
- Medical debt
- Average debt
- Other living quality indicators…

Analytical Techniques

To achieve robust and reliable predictions, we will use a combination of learning models and data analysis:

Supervised Learning Models:

- We will implement various regression models, including Linear Regression, CART (Classification and Regression Trees), and Random Forests, to predict homelessness rates based on socioeconomic and demographic features.
- We will also explore Boosting Techniques (e.g., Gradient Boosting) for improved predictive performance by reducing bias and variance.

Exploratory Data Analysis (EDA):

- We will conduct a thorough EDA to visualize trends, correlations, and outliers in the data. This will help in identifying key factors contributing to changes in homelessness rates.

Model Evaluation:

- We will evaluate model performance on a holdout validation set using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure forecasting accuracy.
- For classification tasks (e.g., predicting high-risk regions), we may use additional metrics such as precision, recall, and F1 score to assess the model's effectiveness in identifying at-risk communities.

Hyperparameter Tuning:

- We plan to test different models and select the best-performing one based on initial evaluation metrics.
- The selected model will undergo hyperparameter tuning using Grid Search to optimize its performance. We will fine-tune parameters such as the number of trees (for Random Forests), learning rate (for Boosting), and maximum depth (for CART).

Impact and Overall Goal

The immediate goal of the project is to predict homelessness in California by counties and have a comprehensive understanding of factors that contribute to the problem. The broader goal is to be able to use this model across the entire United States to better assist public services like social service programs and shelters.

Question:

This multidimensional dataset has counties all across the U.S and one count of homelessness each year for the years 2007 through 2024. What our team was debating about was whether we should predict based on county, how many people are homeless in each county based on all our features (population per density, temperature, housing prices, average years of education, crime… etc) or based on years which would be like training on 2005-2018 for counties in just California and seeing if that model is generalizable across US counties. We are leaning towards the latter but would love any advice on this.