Mitigating Homelessness in the United States through Machine Learning

Sage Wang

Received August 18, 2024
Accepted October 27, 2024
Electronic access November 15, 2024

The growing homelessness crisis in the United States poses a complex challenge that demands innovative approaches. This study seeks to address gaps in traditional models, which have largely concentrated on individual risk factors, by utilizing newly available big data sources. It introduces a novel state-level risk assessment score, identifying 30 key socioeconomic factors correlated with homelessness, drawn from comprehensive nationwide datasets on housing characteristics and shelter inventories. Other key contributions include the development of four robust predictive models, each targeting a specific homeless demographic across four different studies, driven by the need to tailor interventions to the unique challenges and risk factors faced by different population groups. Study 1 achieved an accuracy of 94% and an F1 score of 0.93 when predicting whether homelessness had a greater impact on families with versus without children. Study 2 achieved an RMSE of 1,663 and a R2 value of 0.89 when predicting the total number of homeless individuals by U.S. state. Study 3 achieved an RMSE of 0.096 and an R2 value of 0.388 when predicting the ratio of the number of homeless females to the number of homeless males. Study 4 achieved RMSE values of 1,141, 137, and 1,447, and R2 values of 0.47, 0.86, and 0.82 when predicting the number of homeless individuals in 3 distinct age groups. By leveraging comprehensive datasets and advanced feature selection techniques, this paper provides insights into the broader dynamics of homelessness, facilitating informed decision-making for effective resource allocation and preventive measures.

Introduction

The United States is currently facing a rising homelessness crisis, which has significant social and economic implications. According to the U.S. Department of Housing and Urban Development (HUD), over 650,000 individuals experienced homelessness on a single night in January 2023 [1], the highest this number has ever been. Homelessness continues to rise alongside the cost of living, as illustrated in Figures 1 and 2. Homelessness has been a persistent issue for decades, with patterns closely linked to economic downturns and public health crises. For example, after the 2008 financial crisis, many low-income households faced heightened housing instability as job losses and wage cuts surged across the nation. Similarly, the COVID-19 pandemic exacerbated homelessness, especially as eviction moratoria and emergency financial protections expired. [2] Previous models and approaches, as outlined in the literature review, have primarily focused on individual risk factors—such as childhood adversity and substance abuse—while neglecting broader systemic influences on homelessness at the state level, largely due to the lack of large-scale homelessness data until now. This narrow focus has created a critical gap in the literature: the need for macro-level analyses to guide more comprehensive policy interventions. To address this, our research draws on nationwide datasets from sources like the U.S. Census Bureau and the U.S. Department of Housing and Urban Development.

The impact of the COVID-19 pandemic prompted renewed efforts to improve the assessment of homelessness across the U.S. [3]. With datasets spanning financial, housing, and demographic factors—including DP04, PIT Counts, and Housing Inventory Counts—there is now a unique opportunity to apply interpretable machine learning methods to analyze homelessness at a nationwide scale. This study aims to establish a state-level framework for understanding and predicting homelessness by leveraging multi-source, urban big data. By generalizing findings previously confined to localized levels, it enables the formulation of data-driven policy recommendations at both local and national scales. The research focuses on larger systemic factors, such as overcrowded housing conditions and participation in homelessness mitigation programs, to develop effective state-wide solutions to homelessness.

To accomplish this, this study introduces machine learning models that enable targeted and effective intervention strategies. The sub-objectives include:

1. Development of a State-Level Homelessness Risk Assessment Score: This paper introduces a novel risk assessment score at the state level that incorporates various housing and demographic variables to predict future trends in homelessness rates. This score offers a comprehensive, macro-level tool for policymakers to identify at-risk areas and allocate resources within a state more effectively.
2. Implementation of Targeted Predictive Models: Four distinct

predictive models were developed, each focusing on a specific demographic group within the homeless population. These models provide accurate forecasts for homelessness in families with and without children, total homeless counts, gender ratios, and age group distributions, enabling more targeted and effective intervention strategies.

3. Advanced Feature Selection Techniques: This research employs a variety of feature selection methods, including PCA, comparing p-values, and ANOVA, to identify the most significant factors contributing to homelessness. This approach not only improves the accuracy of the models but also provides transparent, actionable insights for policy interventions.

Literature Review

Traditional methods to combat homelessness have focused on immediate relief efforts, such as charity-run shelters and soup kitchens. With the advance of technology, however, statistical and machine learning approaches have been poised to uncover patterns and predictors of homelessness that may not be evident to the naked eye. Recent studies have employed various methodologies to uncover significant predictors of homelessness, progressively building on each other's findings to enhance the effectiveness of intervention strategies. Doran et al. initiated this by developing a statistics focused screening tool from patient questionnaires to identify emergency department patients at risk of entering a homeless shelter.[4] The resulting screening tool demonstrated an 83.0% sensitivity and a 20.4% positive predictive value for predicting future shelter entry within six months. However, while the study effectively targeted individuals, it did not address broader systemic factors influencing homelessness, a gap that this research aims to fill by focusing on state-wide predictors.

Shelton et al. expanded upon the previous paper by utilizing data from the National Longitudinal Study of Adolescent Health, which surveyed students from 134 high schools, to analyze factors contributing to homelessness among young adults.[5] By applying logistic regression analysis, the study identified key predictors such as childhood adversity, socioeconomic disadvantage, and mental health problems. However, the study was limited to a statistical correlation analysis, and did not report any classification prediction results. This motivated the design of Study 1 in this research, which explores the severity of homelessness in families with versus without children and provides insights into the importance of early intervention efforts. One limitation of Shelton et al.'s study, though, is its limited focus on a specific demographic: high school students. The studies conducted in this paper extend Shelton et al.'s findings by considering a variety of demographic factors such as age, gender, and family status, allowing for a more comprehensive analysis that is not limited to young adults.

Shinn et al. takes this a step further by developing models to determine which families would most benefit from homelessness prevention services in urban areas such as New York City.[6] By utilizing Cox regression to analyze administrative records, Shinn et al. identified key predictors of shelter entry, such as employment status and eviction threats, helping to efficiently target families who are most likely to benefit from these services. This motivated Study 4's multi-regression approach to predicting homeless counts by age group, allowing for deeper insights into such demographic factors. However, Shinn et al.'s approach is limited by its geographic specificity to urban contexts. In contrast, the studies discussed in this paper analyze homelessness across all U.S. states, allowing for a more comprehensive understanding of the factors driving homelessness in not only in urban centers but also in suburban and rural areas. By building on these studies, this paper integrates diverse datasets, models, and feature selection techniques to identify general trends for the broader state population, providing insights for policymakers at a bird's eye view.

Results

A. State-wide Risk Assessment Score

To provide policymakers and researchers with a clear and intuitive way to visualize national homelessness trends, this paper introduces a state-wide homelessness risk score ranging from -5 to 5. This score is based on homelessness counts from the PIT Counts By State dataset and serves as an easily interpretable measure of change. A score of -5 represents a 100% decrease in the number of homeless individuals from 2022 to 2023, while a score of 5 reflects a 100% increase. The majority of states (94%) experienced changes in homelessness rates between -50% and 50%, with only 6% as outliers, showing variations from -60% to 60%. This scale was specifically designed to allow for effective color differentiation on a U.S. heatmap, making it easier to identify trends without overwhelming the audience with excessive detail. Figure 3 displays all U.S. states colored according to their risk scores. Currently, the scores for all U.S. states range between -3 and 3, with the values -5, -4, 4, 5 reserved for potential future, more drastic changes in homelessness rates. While it is highly unlikely that a state will experience a change beyond a 100% increase or decrease in homelessness, given that such a value implies either doubling or completely removing the homeless population, it is recommended for future research to clip the percent change to -100% or 100% if this occurs. The goal of our research is to create mechanisms for policymakers to make informed decisions and drive homelessness to 0 people. As such, this risk score provides an intuitive means for non-technical professionals to view the trends of homelessness across the nation: a ceiling of 100% increase will capture trends of a large increase whereas a 100%

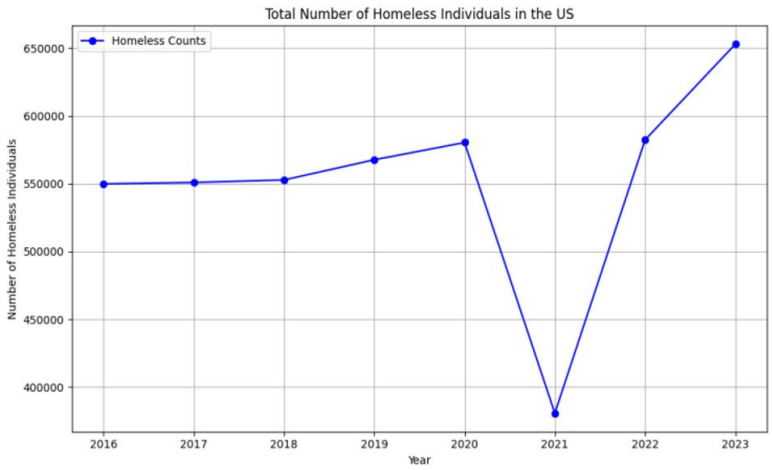

Fig. 1
Homelessness Growth Over Time

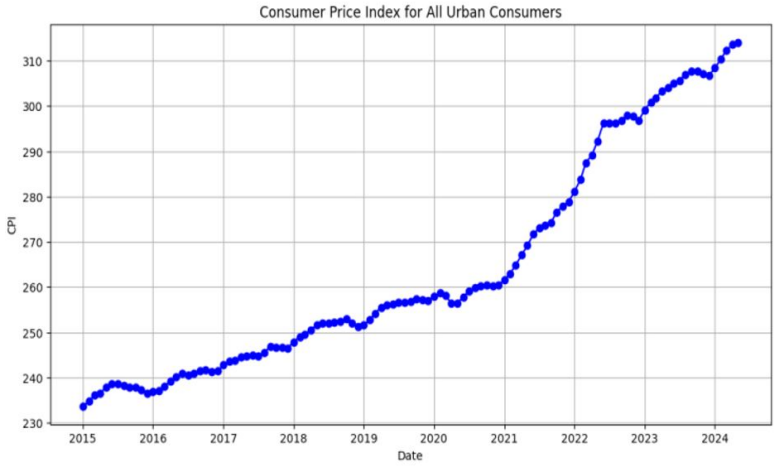

Fig. 2
Consumer Price Index Over Time

decrease will capture the elimination of the homeless population within a region.

Tables A.1 and A.2 present significant features selected by Analysis of Variance (ANOVA) from the HIC and DP04 datasets, respectively. An asterisk (*) next to the feature name means the feature has a statistically significant p-value of less than or equal to 0.05. Analysis of these tables reveals that one of the most prominent features is the HMIS (Homeless Management Information System) participation rate for year-round emergency shelter beds. This suggests that the ability to track and respond to homelessness through coordinated data systems plays a crucial role in reducing homelessness risk, underscoring the importance of robust data management and interagency collaboration. Another critical feature identified is the number of rapid re-housing (RRH) beds available for households with only children. The wide variation in the availability of this value across states with different risk scores highlights disparities in investment and resource allocation. Similarly, the availability of emergency shelter units for households with children also emerged as an important factor. States that have invested in these units are likely better positioned to provide immediate shelter to vulnerable families, thereby reducing their overall homelessness risk. The use of alternative heating fuels and the prevalence of older housing stock also emerged as significant. Reliance on older heating fuels like wood indicates economic hardship in certain regions, which could contribute to a higher risk of homelessness. Additionally, areas with a higher concentration of housing units built before 1939 may struggle with housing disrepair and poor maintenance. On the other hand, the presence of large apartment complexes, as indicated by the feature “UNITS IN STRUCTURE - 20 or more units,” reflects dense urban areas where housing affordability is

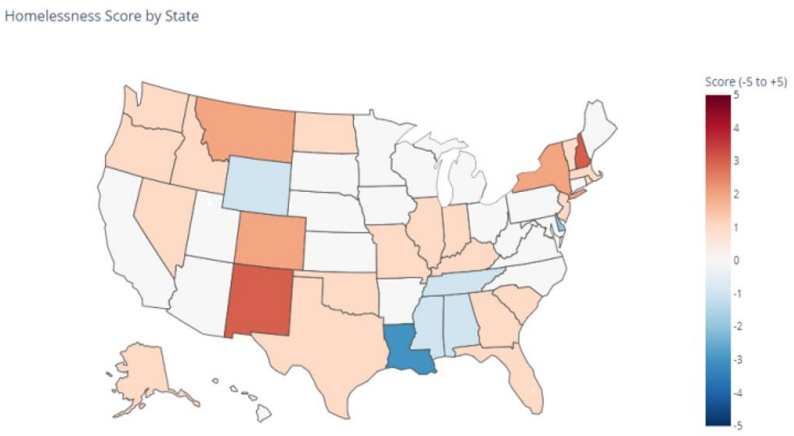

Fig. 3
State-Wide Homelessness Risk Score

a concern, particularly if median rents are high.

Analyzing the trends for the values of these features and comparing them to their corresponding risk scores enables policymakers to make informed decisions on strategic planning and resource distribution for their state. For instance, increasing HMIS participation, renovating old housing units, and building new emergency shelter units could all be key strategies to reduce homelessness. Increasing HMIS participation enables policymakers to track homeless individuals more effectively, optimizing resource distribution and connecting people with essential services. Renovating old housing units expands the availability of affordable housing, improves safety, and reduces financial burdens for at-risk populations. Building new emergency shelter units provides immediate relief to those with physical or mental health challenges, offering critical support and a pathway to stable housing through HMIS integration.

B. Study 1: Comparing With/Without Children

The goal of this experiment is to predict whether there will be more homelessness involving households with or without children using the Housing Inventory Count (HIC) features. This task is important because it sheds light on how to implement targeted interventions to support vulnerable populations such as families with children. The binary target variable was created by comparing two specific columns in the dataset: "Total Beds for Households with Children (ES, TH, SH)" and "Total Beds for Households without Children (ES, TH, SH)." If the value in the first column is less than the value in the second column for a particular row, the target value is set to 1, indicating higher homelessness without children, and 0 otherwise. These two columns were then removed from the dataset to prevent leakage of the target variable into the model features.

Multiple models were employed to achieve the best predictive performance. The models included Random Forest, AdaBoost, and Logistic Regression. After Logistic Regression outperformed its competitors, several types of feature selection were tested with Logistic Regression, including PCA, statistically significant p-values, and highest correlation coefficients. PCA was tested because dimensionality reduction is crucial to managing complexity in a dataset with 77 features. Out of the different feature selection methods tested, selecting features by their p-values performed the best.

The GridSearchCV hyperparameters chosen for this model are displayed in Table B.1. C represents the inverse of the regularization strength, where smaller values specify stronger regularization. Penalty specifies the norm of the penalty. There was a notable class imbalance in the dataset, with a larger proportion of rows having higher homelessness with children. The weighted average metrics (precision, recall, F1-score) accounted for this imbalance, ensuring the evaluation was fair and representative of both classes. This is crucial in a scenario where misclassification could lead to inadequate support for vulnerable populations. The results of all models are summarized in Table 1 and the confusion matrix is displayed in Figure 4.

In Figure 4, the cell where the ground truth label is "Without Children" and the predicted label is "With Children" represents two false positives. This indicates that the model occasionally misclassifies households without children as having a lower likelihood of homelessness. Notably, however, the confusion matrix shows no false negatives, meaning the model does not incorrectly predict higher homelessness involving children as higher without children. Despite the class imbalance, the best performing model still achieved an F1-score of 0.93, indicating that the model can be trusted to make reliable predictions even when the distribution of target labels is skewed. This reveals that the model is particularly reliable in identifying situations where

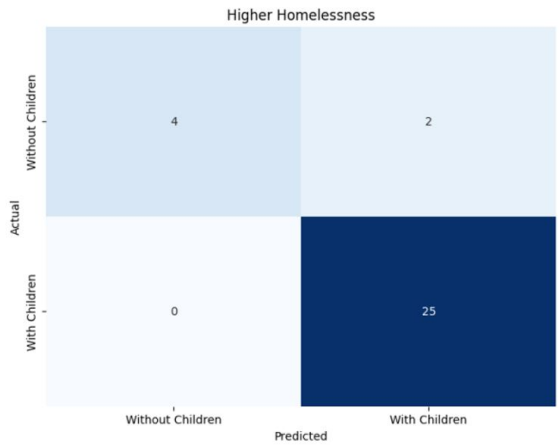

Fig. 4
Study 1: Confusion Matrix for Logistic Regression (10 Features by Lowest P-Values)

homelessness involving children is higher.

The selected features by their p-values are displayed in Table 2. Features such as "Total Units for Households with Children (ES)" and "Total Beds for Households with Children (ES)" rank highly with very low p-values, indicating a strong statistical significance. This is expected because the availability of emergency shelter (ES) units and beds directly impacts homelessness rates. Families with children are particularly vulnerable as they require not just space, but also stability and safety for their children's well-being. Based on these findings, policymakers should prioritize the expansion of emergency shelter capacity specifically targeted for families with children, ensuring they are equipped with safety and support systems necessary for children's well-being. Furthermore, the presence of features related to Homeless Management Information System (HMIS) participation rates suggest that effective management and reporting across a state plays a crucial role in addressing homelessness.

The logistic regression coefficients in Table 3 further support the importance of targeted interventions for households with children. Negative coefficients, such as for "Total Units for Households with Children (ES)"(-2.63) and "Total Beds for Households with Children (ES)"(-1.98), suggest that increasing shelter resources specifically for families with children decreases the likelihood of higher homelessness among households without children. This may be because when families with children receive adequate shelter and support, there is less overall strain on the general shelter system, reducing competition for available resources among all households, including those without children. Meanwhile, positive coefficients like "Total HMIS Year-Round Beds (ES, TH, SH)"(1.45) indicate that increases in general shelter beds may be associated with higher homelessness without children, possibly reflecting a mismatch in resource allocation. These findings underscore the need for policies that expand targeted shelter capacity for families with children, which not only supports this vulnerable group but also contributes to a more stable overall system for addressing homelessness.

C. Study 2: Predicting Total Number of Homeless

The purpose of this experiment is to predict the total number of homeless individuals (based on the PIT Counts By State) using the DP04 Housing Characteristics. Again, various feature selection methods were tested to focus on the most influential features in the dataset. Table 4 demonstrates that the best performing model was Random Forest using features selected by highest correlation coefficients, with a RMSE of 1,663 people and an R2 value of 0.89. The hyperparameters selected by GridSearchCV are shown in Table B.2. n_estimators is the total number of decision trees in the forest. max_depth represents the maximum depth of each tree. min_samples_split controls the minimum number of samples required to split a node of a tree.

The ground truth versus predicted values have been visualized in Figure 5. Considering the complexity of factors that contribute to homelessness, the model provides a fairly accurate prediction of homelessness across the United States. Deviating from the ground truth homeless counts by around 1,600 people is reasonable, as this represents a small error of approximately 0.94% when considering the wide variability in homeless populations across different states, ranging from several hundred in Wyoming to over 170,000 in California.

The top 10 features selected by highest correlation coefficients are shown in Table 5. Many of these features relate to housing costs and room occupancy, which are logical predictors of homelessness. For example, high gross rent and high-value properties indicate areas with high living costs, likely contributing to housing insecurity. Similarly, higher

Table 1: Classification Metrics Model Comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (10 Features by Lowest P-Values)	**0.94**	**0.94**	**0.94**	**0.93**
Logistic Regression (10 Features by Highest Correlation)	0.9	0.9	0.9	0.9
Logistic Regression	0.9	0.9	0.9	0.9
Logistic Regression (10 PCA Components)	0.87	0.89	0.87	0.88
Random Forest	0.77	0.65	0.77	0.7
Adaboost	0.74	0.71	0.74	0.72

Model Results for Predicting Homelessness With vs. Without Children (Study 1)

Table 2: Features Compared By P-Value

Feature	P Value
Total Units for Households with Children (ES)	0.0025
Total Beds for Households with Children (ES)	0.0035
Total Units for Households with Children (ES, TH, SH)	0.0077
HMIS Participation Rate for Year-Round Beds (TH)	0.0285
Total HMIS Year-Round Beds (ES)	0.0428
Total Non-DV Year-Round Beds (ES)	0.0521
Total Year-Round Beds (ES)	0.0626
Total HMIS Year-Round Beds (ES, TH, SH)	0.077
Total Overflow Beds (ES)	0.0795
Total Beds for Households with only Children (ES, TH, SH)	0.1

Top 10 Features by P-Value for Predicting With vs. Without Children (Study 1)

Table 3: Logistic Regression Coefficients

Feature	Coefficient
Total Units for Households with Children (ES)	-2.63
HMIS Participation Rate for Year-Round Beds (TH)	-2.25
Total Beds for Households with Children (ES)	-1.98
Total Units for Households with Children (ES, TH, SH)	-1.97
Total HMIS Year-Round Beds (ES, TH, SH)	1.45
Total HMIS Year-Round Beds (ES)	1.26
Total Beds for Households with only Children (ES, TH, SH)	0.87
Total Year-Round Beds (ES)	0.87
Total Overflow Beds (ES)	-0.77
Total Non-DV Year-Round Beds (ES)	0.58

Logistic Regression Coefficients Using 10 Features by Lowest P-Values (Study 1)

occupancy per room suggests overcrowded living conditions, which could be a precursor to homelessness. Features like “No bedroom” or “1 room” may indicate marginal housing situations like small, overcrowded spaces that risk stepping into homelessness if circumstances worsen. The Random Forest Gini Importance rankings in Table 6 further support the significance of these housing characteristics. The features with the highest coefficients being "ROOMS - 1 room"and "GROSS RENT - $3,000 or more"once again draw attention to housing affordability and space constraints as stressors that contribute to homelessness. These identified features provide actionable insights for policymakers. For instance, areas within a state that struggle with high gross rent or low room counts should be targeted for interventions. The strong predictive performance of the Random Forest model could also enable policymakers to forecast homelessness trends, enabling more proactive resource allocation.

D. Study 3: Gender Ratio Prediction

The goal of this experiment is to predict the ratio of the number of homeless females to the number of homeless males. Once again, the data for the features came from the DP04 Housing Characteristics, and the data for the target column came from the PIT Counts By State. Table 7 shows that the best performing model was Linear Regression with PCA with a RMSE of 0.096 and an R2 value of 0.388. The ground truth versus predicted values have been visualized in Figure 6. Given that the target values range from 0 to 1, an RMSE value of 0.096 represents a relatively low error of 9.6%, indicating a moderate degree of predictive accuracy. The low R2 value is likely due to the lack of gender-specific features in the dataset.

The linear regression coefficients in Table 9 illustrate how the PCA-transformed features relate to predicting the gender ratio of homelessness, though the nature of PCA obscures the relationship between the original features and the combined ones. Components 6 and 9, with the largest negative coefficients, suggest that the underlying features in these components are associated with a lower ratio of homeless women to men, while components 10 and 7, with positive coefficients, indicate the opposite.

Table 8 displays the top 10 features selected by PCA. These features include housing characteristics that offer socioeconomic insights into an area, which can greatly influence the gender distribution of homelessness. For instance, the type of heating fuel used, such as electricity or utility gas, ties into the economic status of households in a region.[7]. Areas that rely on more expensive or less efficient heating methods experience higher living costs, which could disproportionately affect women, who often have lower incomes than men. This economic disparity might lead to a higher rate of homelessness among women in such areas. Additionally, areas with fewer available vehicles might see a higher ratio of homeless women because transportation barriers can make it difficult to maintain stable housing and employment. Finally, population being a top feature is logical because larger populations tend to have more varied social challenges and economic disparities, which can influence homelessness gender ratios. For example, women face greater risks of domestic violence and financial instability, both of which are significant contributors to homelessness. Predicting the gender distribution within homeless populations helps in designing gender-specific support services and policies. For instance, recognizing a higher ratio of homeless women might prompt the development of shelters and services catering specifically to women’s needs. This study aids in ensuring that both men and women receive appropriate assistance.

E. Study 4: Age Group Prediction

The purpose of this experiment is to predict the total number of homeless individuals in each of three age groups provided by the PIT Counts By State: Under 18, 18 to 24, and Over 24. The best model was considered the one with the greatest sum of R2 values, because it represents the model’s overall ability to explain the variance across all age groups. As seen in Figure 7, the distribution of individuals in these groups is unbalanced, with a majority falling into the Over 24 category. Table 10 shows that the best performing model was Random Forest with no feature selection. The model achieved R2 values of 0.47, 0.86, and 0.82 for each respective category while also achieving the smallest average RMSE values of 1,141, 137, and 1,447. This demonstrates a fair level of predictive accuracy. In particular, the R2 values for the 18 to 24 and Over 24 age groups imply the model explains a significant proportion of the variance in the data. The highest RMSE being the Over 24 category is expected given that it has the biggest range of values. The hyperparameters selected by GridSearchCV for this model are displayed in Table B.3.

In a multi-regression task where the goal is to predict outcomes across three different age groups, having access to more information is crucial for capturing the nuances of each category. Avoiding feature selection may have resulted in better performance for this task because the dataset contains a wide range of features that collectively enhance predictive accuracy across all age groups. Retaining the full spectrum of data likely helped in identifying and leveraging complex relationships that were relevant to each specific prediction target.

It is notable that Logistic Regression outperformed Random Forest in predicting the minority classes, namely the Under 18 and 18-24 age groups. This can be attributed to Logistic Regression’s linear nature, which allows it to effectively handle imbalanced data by focusing on fitting the entire dataset, thereby capturing the characteristics of smaller classes more effectively. As a result, Logistic Regression achieved higher accuracy in

Table 4: Regression Metrics Model Comparison

	CORRELATION		MUTUAL INFO		NO FEATURE SELECTION		PCA		P-VALUES	
Model	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**
RandomForest	1663	0.89	2000	0.84	2631	0.72	2644	0.71	2940	0.65
LinearRegression	1713	0.88	4859	0.04	4740	0.08	3462	0.51	7410	1.24
AdaBoost	1795	0.87	2048	0.83	3630	0.46	3553	0.48	3425	0.52

Model Results for Predicting Total Number of Homeless Individuals (Study 2)

Table 5: Features Compared By Correlation Coefficient

Feature	**Correlation Coefficient**
GROSS RENT - $3,000 or more	0.9901
GROSS RENT - $2,500 to $2,999	0.9803
VALUE - $1,000,000 or more	0.9731
GROSS RENT - $2,000 to $2,499	0.9549
SELECTED MONTHLY OWNER COSTS - $3,000 or more	0.9544
OCCUPANTS PER ROOM - 1.51 or more	0.9534
VALUE - $500,000 to $999,999	0.9318
OCCUPANTS PER ROOM - 1.01 to 1.50	0.9194
BEDROOMS - No bedroom	0.9175
ROOMS - 1 room	0.9131

Top 10 Features by Correlation Coefficient for Predicting Total Number of Homeless (Study 2)

Table 6: Random Forest Gini Importances

Feature	**Gini**
ROOMS - 1 room	0.214
GROSS RENT - $3,000 or more	0.2134
BEDROOMS - No bedroom	0.1506
VALUE - $500,000 to $999,999	0.1176
GROSS RENT - $2,000 to $2,499	0.074
VALUE - $1,000,000 or more	0.0719
GROSS RENT - $2,500 to $2,999	0.0607
OCCUPANTS PER ROOM - 1.01 to 1.50	0.043
SELECTED MONTHLY OWNER COSTS (SMOC)	0.0373
OCCUPANTS PER ROOM - 1.51 or more	0.0177

Random Forest Gini Importance using Top 10 Features by Highest Correlation Coefficient (Study 2)

Table 7: Regression Metrics Model Comparison

	PCA		CORRELATION		MUTUAL INFO		NO FEATURE SELECTION		P-VALUES	
Model	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**
Linear Regression	0.096	0.388	0.106	0.254	0.11	0.201	0.264	-3.629	0.12	0.053
AdaBoost	0.108	0.232	0.103	0.291	0.105	0.269	0.108	0.222	0.122	0.007
RandomForest	0.112	0.167	0.106	0.25	0.105	0.276	0.113	0.16	0.13	-0.127

Model Results for Gender Ratio Prediction (Study 3)

Table 8: Features Compared by PCA Value

Feature	PCA Value (Sum of Absolute Contributions)
Population	2.4024
HOUSE HEATING FUEL - Electricity	2.1069
HOUSE HEATING FUEL - Utility gas	2.0399
UNITS IN STRUCTURE - 1-unit, detached	1.6908
HOUSE HEATING FUEL - No fuel used	1.6253
HOUSE HEATING FUEL - Fuel oil, kerosene, etc.	1.5365
YEAR STRUCTURE BUILT - Built 1939 or earlier	1.3899
UNITS IN STRUCTURE - 1-unit, attached	1.3335
VEHICLES AVAILABLE - No vehicles available	1.2632
VALUE - $500,000 to $999,999	1.2379

Top 10 Features Selected by PCA for Gender Ratio Prediction (Study 3)

Table 9: Linear Regression Coefficients

Principal Component	Coefficient
Component 1	-7.74E-10
Component 2	-9.92E-09
Component 3	-4.77E-08
Component 4	-1.82E-09
Component 5	3.24E-08
Component 6	-1.03E-07
Component 7	2.47E-08
Component 8	-1.29E-07
Component 9	-1.72E-07
Component 10	1.47E-07

Linear Regression Coefficients Using Top 10 PCA Features (Study 3)

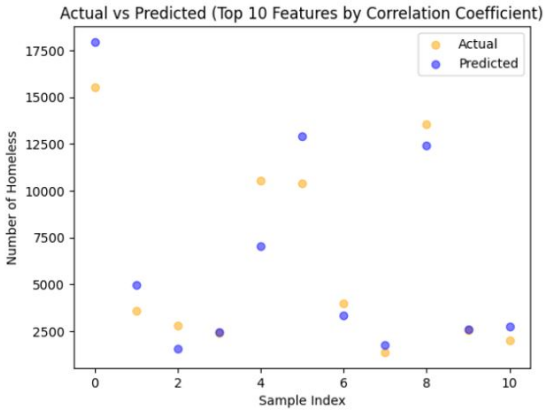

Fig. 5
Study 2: Random Forest Predictions for Total Number of Homeless Individuals

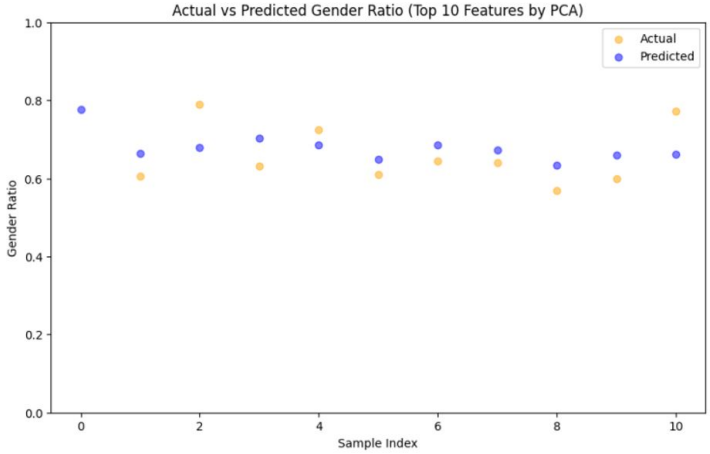

Fig. 6
Study 3: Linear Regression Ground Truth vs. Predicted

predicting the smaller age groups, where the data distribution was more straightforward and less affected by outliers or complex interactions.

On the other hand, Random Forest excels in predicting the majority class, in this case the Over 24 age group, where the data is more abundant but also more variable. Random Forest's ability to model non-linear relationships and handle large datasets with diverse patterns makes it well-suited for predicting this group. However, this complexity can lead to overfitting for minority classes, where the patterns are less pronounced. This explains why Random Forest has a higher RMSE in the minority groups while performing robustly in the majority class. Identifying age groups that are more susceptible to homelessness allows policymakers and social services to tailor interventions more effectively. For example, if there is a predicted increase in the number of homeless individuals over 24 years old, efforts can be concentrated on providing job training and affordable housing to this demographic.

Methodology

A. Datasets

1) Housing Inventory Counts (HIC): The HIC dataset, collected by the U.S. Department of Housing and Urban Development (HUD), provides Housing Inventory Count data by geographic service area and by state.[8]. This dataset was selected for its detailed coverage of housing inventory and service utilization, which is crucial for understanding the capacity of homeless services at a state level. The dataset comprises 77 features, capturing a wide array of information ranging from service utilization to housing inventory specifics. This paper focuses on

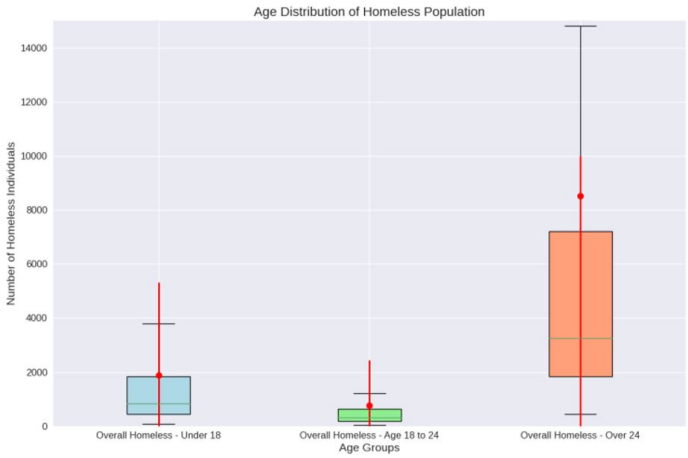

Fig. 7
Study 4: Homeless Distribution by Age Group Across States

the 55 rows of data collected in 2022.

2) Point in Time (PIT) Counts By State: The PIT Count dataset is another valuable resource provided by the HUD. This dataset offers a snapshot of both sheltered and unsheltered individuals experiencing homelessness on a single night in January each year by state, along with estimates of chronic homelessness. [8]. The PIT Count dataset was selected for its granularity in tracking different demographic categories of homelessness, which is essential for analyzing variations in homelessness across states. The dataset contains 55 rows and 575 columns. However, this paper focuses on key columns from 2022, including: Overall Homeless, Overall Homeless - Under 18, Overall Homeless -18 to 24, Overall Homeless - Over 24, Overall Homeless - Female, and Overall Homeless - Male.

3) DP04 Housing Characteristics: The DP04 Housing Characteristics dataset is sourced from the US Census Bureau's American Community Survey. [9] This dataset was compiled by downloading individual data tables for each U.S. state and then concatenating them into a single dataset consisting of 52 rows and 140 columns. The smaller number of rows is due to the absence of data for three U.S. territories in this dataset, which were included in the previous two datasets. To maintain consistency, those 3 territories were removed from the other two datasets. The DP04 dataset provides a comprehensive view of housing characteristics such as the number of housing units, bedroom counts, gross rent, and heating fuel types. This dataset was selected because such detailed characteristics allow for in-depth analysis of their potential correlations with homelessness rates.

B. Metrics

1) MSE: Mean Squared Error (MSE) is a metric used to evaluate the accuracy of regression models. It measures the average of the squared differences between the predicted and actual values. MSE is calculated by taking the average of the squares of the errors:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value, $\hat{y}$ is the predicted value, and n is the number of observations. MSE penalizes larger errors more significantly than smaller ones due to the squaring of the differences, making it sensitive to outliers.

2) RMSE: Root Mean Squared Error (RMSE) is a metric that measures the square root of the Mean Squared Error. It provides a way to measure the magnitude of prediction errors in the same units as the original data. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i is the actual value, $\hat{y}$ is the predicted value, and n is the number of observations. RMSE is particularly useful for the prediction tasks in this research because it gives a sense of how large the prediction errors are, on average, making it more interpretable in the context of the selected homelessness datasets.

3) Accuracy: Accuracy is a metric commonly used to evaluate classification models. It measures the proportion of correct predictions out of the total number of predictions made. Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

It provides a straightforward measure of a model's performance, indicating how often the model correctly predicts the target class.

Table 10: Multi-Regression Metrics Model Comparison

Model	RMSE	R2
NO FEATURE SELECTION		
RandomForest	[1141, 137, 1447]	[0.47, 0.86, 0.82]
AdaBoost	[1252, 169, 2012]	[0.37, 0.78, 0.66]
LinearRegression	[1197, 334, 3630]	[0.42, 0.16, -0.11]
MUTUAL INFO		
RandomForest	[1264, 185, 1860]	[0.36, 0.74, 0.71]
AdaBoost	[1334, 203, 1896]	[0.28, 0.69, 0.70]
LinearRegression	[1157, 338, 4257]	[0.46, 0.14, -0.52]
CORRELATION		
RandomForest	[956, 155, 2583]	[0.63, 0.82, 0.44]
AdaBoost	[814, 165, 3734]	[0.73, 0.79, -0.17]
LinearRegression	[1146, 237, 3433]	[0.47, 0.58, 0.01]
P-VALUES		
RandomForest	[1381, 301, 2354]	[0.23, 0.32, 0.54]
AdaBoost	[1482, 282, 2406]	[0.11, 0.40, 0.51]
LinearRegression	[999, 522, 6731]	[0.60, -1.05, -2.80]
PCA		
RandomForest	[1268, 207, 1536]	[0.35, 0.68, 0.80]
AdaBoost	[1371, 245, 2519]	[0.24, 0.55, 0.47]
LinearRegression	[501, 282, 3429]	[0.90, 0.40, 0.01]

Predict Number of Homeless Individuals in Each Age Group (Study 4)

4) F1 score: The F1 score is a metric used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets. It is the harmonic mean of precision and recall, providing a balance between the two. The F1 score is calculated as:

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision is the ratio of true positive predictions to the total predicted positives, while recall is the ratio of true positive predictions to the actual positives. The F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall. It is useful for scenarios where the cost of false positives and false negatives is high and where a balance between precision and recall is desired.

5) R^2 score: The R^2 score, also known as the coefficient of determination, is a metric used to evaluate the goodness of fit of a regression model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables. The R^2 score is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the actual value, $\hat{y}$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the number of observations. An R2 score of 1 indicates that the model perfectly predicts the dependent variable, while a score of 0 indicates that the model does not explain any of the variability in the data.

C. Feature Selection Techniques

1) Principal Component Analysis: Principal Component Analysis (PCA) is a powerful technique used for dimensionality reduction in machine learning. PCA works by identifying the directions (principal components) along which the variation in the data is maximal, or in other words the most unique information about the dataset is captured. Homelessness data often contains many interrelated variables, which can result in overfitting and computational inefficiencies. PCA addresses this by simplifying the dataset, highlighting the most significant patterns and relationships. This not only enhances interpretability but also allows the analysis to focus on key factors driving data variation. By concentrating on these influential variables, decision-makers can gain clearer insights, essential for developing effective policy interventions.

2) Mutual Information: Mutual Information (MI) quantifies the amount of information obtained about one random variable through another random variable. In the context of feature selection, MI captures both linear and non-linear dependencies between variables, making it highly versatile. Given the complex, non-linear nature of socioeconomic factors contributing to homelessness, MI ensures that no important dependencies are missed. By ranking features based on their MI

scores, the model can focus on the most relevant aspects of the data, leading to better generalization and performance.
3) Analysis of Variance: Analysis of Variance (ANOVA) is a statistical method used to compare the means of three or more groups to determine if there are any statistically significant differences between them. ANOVA was used for this study because of its ability to test the influence of multiple factors simultaneously. However, it assumes that the data is normally distributed and that variances are equal across groups, which might not always hold true in real-world scenarios.
4) T-Test and P-Value: The t-test is a statistical method used to determine if there is a significant difference between the means of two groups, which can be related in certain features. The p-value indicates whether the difference in means is statistically significant, enabling the selection of features that have a tangible impact on homelessness. A low p-value indicates strong evidence against the null hypothesis, suggesting that the feature has a significant impact on the outcome, while a high p-value implies weak evidence against the null hypothesis, indicating that the feature does not have a significant effect.
5) Correlation Coefficient: Correlation analysis is another statistical tool used to measure the strength and direction of the linear relationship between two variables. This technique is particularly useful for identifying straightforward, interpretable relationships, such as the correlation between rent prices and homelessness rates. A correlation coefficient close to 1 implies a strong positive linear relationship, while a correlation coefficient close to -1 indicates a strong negative linear relationship. A correlation coefficient around 0 suggests no linear relationship between the variables.

D. Models

1) Linear Regression: Linear Regression estimates the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Linear Regression aims to minimize the mean squared error of the data points to the trend line:

$$\text{minimize} \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

This model is highly interpretable, making it useful for understanding the direct impact of various factors on homelessness rates. However, its assumption of a linear relationship between variables may not adequately capture the non-linear patterns often present in real-world homelessness data.
2) Logistic Regression: Logistic Regression predicts the probability that an input belongs to a specific class using the logistic function. This probability is represented as:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

This model is straightforward and interpretable, making it suitable for identifying key factors contributing to homelessness. However, like Linear Regression, its linear nature may not capture complex patterns in the data, potentially limiting its effectiveness.
3) AdaBoost: AdaBoost combines multiple weak classifiers to create a strong classifier by focusing on hard-to-classify cases. The weight of a weak classifier at iteration t is represented as:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - TotalError}{TotalError}$$

For predicting homelessness, it can improve accuracy by emphasizing critical features. However, it is sensitive to noisy data and outliers, which are prevalent in real-world homelessness data, potentially reducing its reliability. As a boosting algorithm, AdaBoost has the advantage of progressively learning to focus on more challenging data samples, which are often overlooked by simpler models.
4) Random Forest: Random Forest builds multiple decision trees and aggregates their predictions to enhance accuracy and prevent overfitting. For regression tasks, the final prediction is the average of the prediction made by all the trees:

$$\hat{f} = \frac{1}{N} \sum_{i=1}^{N} f_i(X')$$

Random Forest's robustness is beneficial for handling the diverse and high-dimensional nature of the homelessness datasets being used. However, the model can be computationally expensive and less interpretable, which may complicate the analysis of key factors for policymakers. Nevertheless, the tree-based structure of Random Forests offers a robust decision-making framework, making it a valuable tool for addressing the complexity of homelessness data.

E. Hyperparameter Tuning

GridSearchCV is a technique used to identify the optimal set of hyperparameters that yield the best performance for a given algorithm. GridSearchCV works by exhaustively searching through a specified parameter grid, evaluating every possible combination of hyperparameters through cross-validation. During this process, the model is trained and validated multiple times, each with a different set of hyperparameters. The combination that results in the best score is then selected as the optimal set of hyperparameters.

Conclusion

This research demonstrates the potential of using state-level data to understand and address homelessness more effectively. Four different studies were designed involving varying homeless

populations based on age, gender, and family composition. Machine learning models were chosen over deep neural networks in order to maximize interpretability and transparency, which is crucial in safety-critical domains like homelessness. Historically, machine learning approaches are more trusted in such domains, and while neural networks continue to evolve, they are still working towards achieving the level of reliability and clarity needed for these applications.[10] A statewide homelessness risk score was introduced to quantify and visualize trends. This paper's analysis identified significant features, such as HMIS participation rates and house heating fuel types, which impact these scores. These findings provide a clear framework for policymakers to pinpoint high-risk areas and allocate resources more strategically to reduce homelessness.

Despite the strengths of our approach, there are some limitations to consider. The data for all four studies was collected in 2022, during a time when populations were still recovering from the COVID-19 pandemic, which may have influenced our results. Future research should include data from multiple years to better capture long-term trends. Incorporating a time series analysis could also offer deeper insights into the temporal dynamics of homelessness and improve the models' predictive power.

Overall, this study offers a solid foundation for future research and practical recommendations for addressing homelessness at the state level. By identifying high-risk areas and key contributing factors, policymakers are equipped with tools to design targeted interventions. This data-driven approach is a crucial step toward developing more efficient and equitable solutions to social challenges, paving the way for systematically reducing and preventing homelessness through informed planning and proactive measures.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Samuel Lefcourt, for guiding me through this project, helping me with the research process, and providing his invaluable feedback on the paper.

References

1 *HUD releases January 2023 point-in-time count report*, 2023, https://www.hud.gov/press/press_releases_media_advisories/hud_no_23_278, Accessed: 2024-10-25.

2 M. Ring and J. Schuetz, *Homelessness fell across most metro areas after the Great Recession—Will COVID-19 change that?*, 2021, https://www.brookings.edu/articles/homelessness-fell-across-most-metro-areas-after-the-great-recession-will-covid-19-change-that/.

3 J. Tsai and J. Alarcón, *American Journal of Public Health*, 2022, **112**, 633–637.

4 K. M. Doran, E. Johns, S. Zuiderveen, M. Shinn, K. Dinan, M. Schretzman, L. Gelberg, D. Culhane, D. Shelley and T. Mijanovich, *Health Services Research*, 2022, **57**, 285–293.

5 K. H. Shelton, P. J. Taylor, A. Bonner and M. V. D. Bree, *Psychiatric Services*, 2009, **60**, 465–472.

6 M. Shinn, A. L. Greer, J. Bainbridge, J. Kwon and S. Zuiderveen, *American Journal of Public Health*, 2013, **103**, S324–S330.

7 J. R. Castigliego, S. Alisalad, T. Stasio and L. Stanton, *Inflection point: When heating with gas costs more*, 2021, https://aeclinic.org/publicationpages/2021/01/13/inflection-point-when-heating-with-gas-costs-more.

8 *Point-in-Time (PIT) and Housing Inventory Count (HIC) Data Since 2007*, https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/, Accessed: 2024-10-25.

9 *American Community Survey 5-Year Data (DP04): Selected Housing Characteristics*, https://data.census.gov/table?q=DP04, Accessed: 2024-10-25.

10 C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova and C. Zhong, *CoRR*, 2021, **abs/2103.11251**, year.

APPENDICES

Appendix A

Table A.1: HIC Feature Comparison By Risk Score Using ANOVA

Score	-3	-2	-1	0	1	2	3
HMIS Participation Rate for Year-Round Beds (ES)*	0.22	0.73	0.64 ± 0.04	0.69 ± 0.16	0.75 ± 0.11	0.79 ± 0.19	0.64 ± 0.14
HMIS Participation Rate for Year-Round Beds (PSH)*	0.91	0.67	0.71 ± 0.35	0.85 ± 0.11	0.78 ± 0.15	0.55 ± 0.22	0.64 ± 0.01
HMIS Participation Rate for Year-Round Beds (RRH)*	0.8	0.94	0.94 ± 0.06	0.9 ± 0.09	0.84 ± 0.08	0.84 ± 0.11	0.73 ± 0.08
HMIS Participation Rate for Year-Round Beds (ES, TH, SH)	0.29	0.73	0.68 ± 0.05	0.69 ± 0.14	0.73 ± 0.1	0.68 ± 0.16	0.66 ± 0.21
Total Beds for Households with only Children (RRH)	0	0	9 ± 18	0.32 ± 0.78	1.3 ± 2.23	2.5 ± 5	0.5 ± 0.71
Total Beds for Households with only Children (PSH)	0	0	5 ± 9	0.73 ± 2.21	0.05 ± 0.22	1.25 ± 2.5	0
Total Units for Households with Children (ES)	1469	175	224 ± 197	724 ± 1404	738 ± 907	3223 ± 5625	235 ± 69
Total Beds for Households with Children (ES)	4308	560	652 ± 541	2281 ± 4012	2369 ± 2924	9669 ± 17321	720 ± 150
Total HMIS Year-Round Beds (ES)	1317	721	1147 ± 968	4187 ± 8570	4211 ± 4060	16953 ± 30265	1050 ± 162
Total Non-DV Year-Round Beds (ES)	5626	915	1459 ± 1379	5038 ± 10351	4693 ± 4337	18897 ± 33165	1412 ± 403
Total Units for Households with Children (ES, TH, SH)	1518	272	295 ± 254	1008 ± 1907	1001 ± 1062	3514 ± 5808	364 ± 51
Total Year-Round Beds (ES)	5893	985	1782 ± 1551	5662 ± 10796	5392 ± 4796	19958 ± 34873	1706 ± 620
Total HMIS Year-Round Beds (ES, TH, SH)	1880	1034	1556 ± 1313	5423 ± 10724	5470 ± 4970	17955 ± 31410	1439 ± 105

HIC Top 15 ANOVA Features

Table A.2: DP04 Feature Comparison By Risk Score Using ANOVA

Score	-3	-2	-1	0	1	2	3
HOUSE HEATING FUEL - Other fuel	2317	2610	2588 ± 1844	14509 ± 14734	9840 ± 7479	41587 ± 54160	8843 ± 3039
VEHICLES AVAILABLE - No vehicles available	148577	23783	81986 ± 57074	189078 ± 213071	198337 ± 186383	800625 ± 1266927	33948 ± 13963
HOUSING OCCUPANCY - Rental vacancy rate	7.2	3.2	6.58 ± 1	5 ± 1.22	5 ± 1.67	4.5 ± 0.85	4 ± 1.48
UNITS IN STRUCTURE - 2 units	79192	4561	38544 ± 29161	82238 ± 87907	97741 ± 102025	292070 ± 458582	23442 ± 11452
YEAR STRUCTURE BUILT - Built 1939 or earlier	128921	34845	92219 ± 64769	365754 ± 410831	280420 ± 297529	932652 ± 1412023	79257 ± 57175
HOUSE HEATING FUEL - Fuel oil, kerosene, etc.	1732	35708	3122 ± 3616	104723 ± 195194	65697 ± 143938	428233 ± 737011	112191 ± 157574
ROOMS - Median rooms	5.5	6.2	5.78 ± 0.05	5.72 ± 0.46	5.46 ± 0.46	5.47 ± 0.42	5.45 ± 0.21
UNITS IN STRUCTURE - 20 or more units	101602	23289	85838 ± 80495	268827 ± 414807	314685 ± 408559	836797 ± 1160711	47535 ± 4039
HOUSING OCCUPANCY - Homeowner vacancy rate	1.1	0.5	0.72 ± 0.21	0.79 ± 0.21	0.89 ± 0.27	0.77 ± 0.21	0.7 ± 0.57
GROSS RENT - Median (dollars)	984	1274	944 ± 102	1153 ± 338	1266 ± 279	1383 ± 336	1176 ± 312
HOUSE HEATING FUEL - Wood	7191	2536	17387 ± 11121	42343 ± 41665	27797 ± 24606	54998 ± 38289	40600 ± 13627
SELECTED MONTHLY OWNER COSTS - Median ($)	1472	1629	1392 ± 138	1672 ± 449	1830 ± 442	2021 ± 378	1783 ± 530
HOUSE HEATING FUEL - Bottled, tank, or LP gas	39340	40565	88604 ± 40573	155521 ± 132107	104304 ± 85797	191529 ± 183735	82093 ± 30935
GROSS RENT - Less than $500	61298	7107	48464 ± 31113	64804 ± 56618	57207 ± 47745	117070 ± 160766	15462 ± 8770
BEDROOMS - 5 or more bedrooms	54169	24951	76196 ± 52363	156803 ± 148388	139188 ± 134652	214119 ± 163016	24482 ± 330

DP04 Top 15 ANOVA Features

Appendix B

Table B.1: GridSearchCV Hyperparameters

Name	All Options	Selected Option
C	0.01, 0.1, 1, 10, 100	100
Penalty	11, 12	11

Selected GridSearchCV Hyperparameters for Logistic Regression Using 10 Features by Lowest P-Values (Study 1)

Table B.2: GridSearchCV Hyperparameters

Name	All Options	Selected Option
n_estimators	50, 100, 200	50
max_depth	None, 10, 20	20
min_samples_split	2, 5, 10	2

Selected GridSearchCV Hyperparameters for Random Forest with Correlation Coefficients (Study 2)

Table B.3: GridSearchCV Hyperparameters

Name	All Options	Selected Option
n_estimators	50, 100, 200	50
max_depth	None, 10, 20	10
min_samples_split	2, 5, 10	2

Selected GridSearchCV Parameters for Random Forest (Study 4)