OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Predicting Next-Year Homelessness Levels Using HUD Point-in-Time Data

Jaden Cheung
cheung.jaden@gmail.com

ABSTRACT

This study investigates whether machine learning models can perform better than a simple persistence baseline when used to predict next-year homelessness using the U.S. Department of Housing and Urban Development Point-in-Time dataset. The target variable is the next-year overall homelessness count by Continuum of Care region. All models were evaluated using a chronological split, where earlier years were used for training and later years were used for testing. This study also conducted two different experiments. Experiment 1 included the current-year overall homelessness and other count variables as features, but in Experiment 2, those features were removed, leaving only a few features left. In Experiment 1, the linear regression model did slightly better in RMSE than the persistence baseline, but their difference was very small. The persistence baseline also had a lower MAE and MAPE than the linear regression. The baseline's MAE and MAPE was 112.58 and 0.145, while the linear regression's MAE and MAPE was 138.76 and 0.232. In Experiment 2, the persistence baseline did much better than all machine learning models. Therefore, these results suggest that predicting next year's count as the current year's count is a strong predictive signal in the PIT dataset. These models may need more external variables to ultimately perform better than the simple persistence baseline.

INTRODUCTION

Homelessness is one of the major challenges in the United States, as it affects hundreds of thousands of people each year. According to the U.S. Department of Housing and Urban Development, over 650,000 people experienced homelessness in a single night in 2023, showing the scale of this issue. (U.S. Department of Housing and Urban Development, 2023) Homelessness doesn't just impact the people; it uses significant public resources like healthcare and government. Researchers have also shown the importance of moving towards a prevention-based approach to combat homelessness with decisions driven by data. (Culhane et al., 2011) There have been many efforts to predict homelessness levels. (Montgomery et al., 2013; Byrne et al., 2018) However, because of multiple factors, like housing, economic conditions, and different demographics, homelessness levels are difficult to predict. (Desmond, 2016) Studies have shown that homelessness is influenced by multiple variables, which make getting accurate predictions much more difficult and complex. (Greenberg & Rosenheck, 2010) According to

data collected by the U.S. Department of Housing and Urban Development, homelessness is measured annually through the Point-in-Time count. (U.S. Department of Housing and Urban Development, n.d.) This creates a picture of homelessness across different geographic regions at a certain point. Each data point refers to a specific place and year with information like subpopulations and chronically homeless individuals, allowing for an analysis with multiple factors. This analysis discovers patterns, which become a valuable resource for understanding and predicting homelessness. Moreover, machine learning models have the ability to find relationships between factors and make predictions based on datasets. (VanBerlo et al., 2021) In homelessness, these models can show patterns that can allow healthcare and governments to allocate their resources effectively and implement strategies. (Toros et al., 2018)

This study asks whether machine learning models using HUD PIT features can improve next-year homelessness predictions beyond the simple persistence baseline that predicts next year's homelessness count as equal to the current year's count. Instead of seeing which model performs better, the study shows whether complex models add more predictive value beyond year-to-year stability. The main contribution of this study is showing that with PIT-based one-year homelessness features, it is difficult to beat the benchmark of the simple persistence baseline especially when the current-year count variables are removed.

MATERIALS AND METHODS

The dataset was obtained from Kaggle's HUD PIT homelessness dataset.

The dataset used by this study is based on the U.S. Department of Housing and Urban Development Point-in-Time homelessness count. It gives exact data on homelessness across different demographics annually. Every row represents a specific place, a Continuum of Care, and a year. The columns are full of different information, like total homelessness, sheltered and unsheltered populations, and other subpopulations. The primary columns used in this study are overall homelessness count, unsheltered homelessness count, sheltered homelessness count, emergency shelter homelessness count, transitional housing homelessness count, chronically homeless count, homeless veterans count, and year count. These metrics together cover the overall structure of homelessness in certain regions. This study is a forecasting task, the dataset in this model is split chronologically instead of randomly. The dataset contains annual PIT data from 2012 to 2018. 2018 was not used because after creating the next-year variable, the only usable years were 2012 to 2017. Therefore, the training data set is all the data from 2012-2015, the validation data set is from 2016, and the testing data set is from 2017. This split better represents the ability to predict the future homelessness counts and avoids training on future years. The model trains itself using the training data set. Once the model is done training, it tests itself on the testing data set. All of the evaluations and metrics are based on the testing data results. The reason for splitting the data set is to prevent overfitting and to see how the model will do on new data, so there are no biases. There are also many other ways to split a data set. There can also be validation sets to train the model's hyperparameters.

Experiment Design

Two experiments were used to test whether machine learning models could improve predictions beyond a simple persistence baseline. The persistence baseline predicted next-year overall homelessness as equal to current-year overall homelessness. In Experiment 1, all of the PIT features were used including current-year overall homelessness and related count variables. This tests whether the machine learning models could improve the prediction when they could use the current-year homelessness counts. In Experiment 2, the current-year total homelessness and other direct component count variables were removed. The reduced model kept chronically homeless count, homeless veterans count, and count year. Therefore, this would test whether the remaining PIT features could predict the next-year homelessness without relying on last year's count.

Data Preprocessing:

The dataset was sorted by region (CoC_Number) and (Count_Year). The target variable was next-year's overall homelessness count, which we created by shifting the overall homelessness count forward by one year in each region, so we could use last year's count to predict next year's count. Rows with missing values were removed from the dataset

Feature and Target:

The input feature matrix (X) consisted of the following variables: overall homelessness count, unsheltered homelessness count, sheltered homelessness count, emergency shelter homelessness count, transitional housing homelessness count, chronically homeless count, homeless veterans count, and count year. The target variable (y) was next-year overall homelessness count

Linear Regression:

A linear regression model shows the relationships between an independent and dependent variable by drawing a linear relationship between the two variables, so that as the input changes, the output will change at the same constant rate. (James et al., 2021) Linear regression was included as a simple benchmark machine learning model to see how a weighted linear combination of PIT features could perform better than the persistence baseline. This model was useful since its coefficients can be interpreted directly, showing exactly how each feature is associated with next-year homelessness counts.

Decision Tree

The decision tree model makes a prediction by splitting all the data into small groups based on the features. (Quinlan, 1986) A decision tree model was included to test if nonlinear patterns in PIT features could predict next-year homelessness. The model's hyperparameters were tuned on the validation set to reduce overfitting.

Random Forest

A random forest model is an ensemble method that combines multiple decision trees to make a prediction. (Breiman, 2001) A random forest model was included to test if multiple decision trees could find

nonlinear patterns that a single decision tree might not find. The model used 189 decision trees with a maximum depth of 5.

Neural Networks

A neural network was included to test if a more flexible model could have a better performance than a linear regression or decision tree model. Features were standardized before training, and the model had two hidden layers with 64 neurons each.

RNN/LSTM

SimpleRNN and LSTM models were included as comparison models. The dataset does not include multi-year sequences for each region, so the results shouldn't be used as a test of recurrent predictions.

Model Tuning and Evaluation Metrics

All of the models were evaluated using mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R^2. MAE measures the average absolute difference between the predicted and actual counts, RMSE measures the square root of the prediction errors squared, and MAPE measures the average prediction error as a percentage of the actual value. Lower MAE, RMSE, and MAPE mean that the model is performing well, while a higher R^2 value also means that the model is strong.

Table 1a:

Model performance for next-year homelessness prediction using HUD PIT data. Models are evaluated on the chronological test set using mean absolute error, root mean squared error, mean absolute percentage error, and R^2.

Experiment 1: Full PIT Features

Includes current-year overall homelessness and related count variables.

Model	MAE	RMSE	MAPE	R^2
Linear Regression	138.76	**233.06**	0.232	0.997666
Persistence Baseline	**112.58**	233.26	**0.145**	0.997662
Neural Network	323.39	405.69	0.812	0.992928

Model	MAE	RMSE	MAPE	R^2
Weighted Ensemble	228.90	671.72	0.458	0.980612
Decision Tree	641.27	1,382.64	1.544	0.917858
Random Forest	253.70	1,732.25	0.334	0.871065
Ensemble Average	491.88	1,920.83	0.413	0.841465
SimpleRNN Exploratory	1,395.92	5,021.98	1.001	-0.083675
LSTM Exploratory	1,395.78	5,022.04	1.000	-0.083702

Table 1b:

Experiment 2: Reduced PIT Features

Removes current-year total homelessness and direct component count variables.

Model	MAE	RMSE	MAPE	R^2
Persistence Baseline	**112.58**	**233.26**	**0.145**	0.997662
Decision Tree	719.28	1,242.63	1.812	0.933651

Model	MAE	RMSE	MAPE	R^2
Weighted Ensemble	589.10	2,341.92	1.134	0.764336
Random Forest	469.60	2,558.80	0.691	0.718666
Ensemble Average	617.59	3,007.53	0.836	0.611339
Linear Regression	576.56	3,326.86	0.587	0.524425
Neural Network	1,100.14	3,349.67	2.859	0.517882
LSTM Exploratory	1,395.39	5,021.92	0.999	-0.083650
SimpleRNN Exploratory	1,396.17	5,022.14	1.001	-0.083747

Figure 1. Predicted versus actual next-year homelessness counts for the persistence baseline.
Each point represents a Continuum of Care region in the test set. The diagonal line shows the perfect prediction, and the points closely follow this line showing that the persistence baseline performed strongly because the homelessness counts are highly stable from year to year.

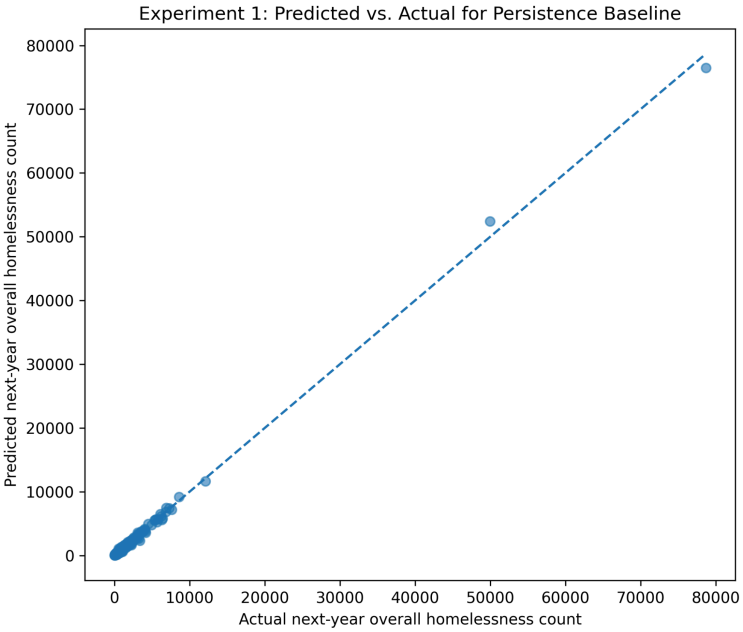
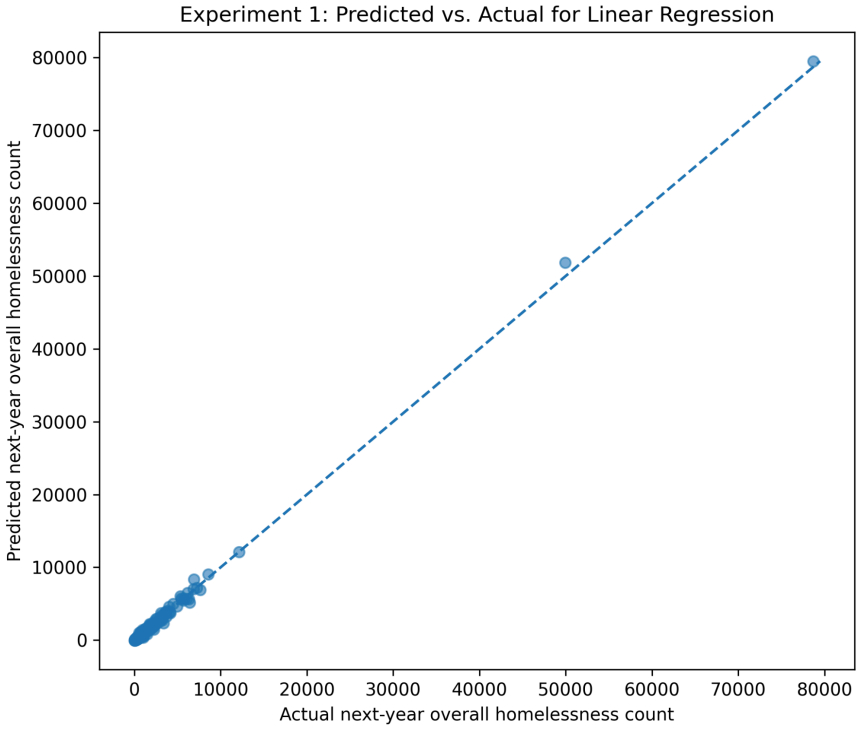


Figure 2. Predicted versus actual next-year homelessness counts for linear regression.
Each point represents a Continuum of Care region in the test set. Linear regression performed the best as a non-baseline model in Experiment 1 by RMSE, but its predictions are very close to the persistence baseline. This shows how the machine learning model provides only a small improvement over the simple baseline when including current-year count variables.



RESULTS

Figures 1 and 2 both show that the persistence baseline and linear regression predictions are very aligned with the actual values, which shows that both the simple models have a strong performance.

The results show that a simple model performed the best in the prediction of next-year homelessness counts. In Experiment 1, which included current-year homelessness count variables, linear regression had the lowest RMSE value at 233.06, while the persistence baseline model has an RMSE of 233.26. This difference was extremely small, and the persistence baseline has a lower value for both MAE and MAPE. In Experiment 2, where current-year total homelessness and other count variables were removed, the persistence baseline outperformed all other machine learning models. The best non-baseline model was the decision tree with an RMSE of 1242.63 and R^2 of 0.933651. The persistence baseline's model has an RMSE of 233.26 and R^2 of 0.997662. These results show that the best way to predict next-year homelessness count in the PIT dataset is to use a year-to-year stability model instead of the complex models.

Table 2: Linear regression coefficients for Experiment 2 reduced PIT features.

Feature	Coefficient	Standard Error	95% CI Lower	95% CI Upper
Intercept	-156346.22	109850.67	-371653.52	58961.08
Chronically homeless count	2.03	0.19	1.66	2.41
Homeless veterans count	7.48	0.42	6.66	8.31
Count year	77.60	54.56	-29.33	184.53

The reduced-features linear regression model shows positive coefficients for chronically homeless count and homeless veterans count, which means the more people in those categories, the more people in next-year overall homelessness. However, the linear regression model still performed much worse than the persistence baseline. This means that these features have some predictive signal, but not enough to beat year-to-year stability.

DISCUSSION

The results show that a simple baseline is necessary when evaluating prediction models. Since homelessness counts are stable from year to year, the persistence baseline model performs very well. For example, in Experiment 1, linear regression did slightly better than the baseline, but the difference was very small. On the other hand, the baseline had much lower scores in MAE and MAPE. This shows that even when machine learning models perform well, they still might not do better than a simple persistence baseline.

Experiment 2 further supports this. When current-year total homelessness and the other direct count variables were removed, all machine learning models performed worse than the persistence baseline. This shows that the reduced features like chronically homeless count and homeless veterans count did not contain enough information to beat the persistence baseline. This conclusion shows the challenge of predicting homelessness as it is influenced by multiple factors beyond PIT counts. Some factors that

influence homelessness could be housing costs, economic conditions, and local policy. Future work would be adding external features such as unemployment, poverty rates, housing supply, and local housing policy. Housing cost features may be especially important as cost of living has been directly associated with homelessness levels (Heston, 2023). Future research should also compare models carefully rather than just saying that a more complex model is better. In this study, the persistence baseline was very difficult to beat, showing how homelessness needs to be evaluated against simple and appropriate baselines (Messier et al., 2021).

CONCLUSION

This study shows that simple prediction baselines are essential when evaluating machine learning models for predictions around homelessness. In Experiment 1, linear regression had a slightly improved value of RMSE over the persistence baseline. However, in Experiment 2, where a lot of the features were removed, the persistence baseline easily outperformed all other machine learning models. This brings us to the conclusion that with HUD PIT data, the best prediction for next year homelessness levels is the current year homelessness levels. More complex models did not add any predictive value unless they had all the features in the dataset. Therefore, the persistence baseline should be used as the benchmark for future homelessness prediction models using PIT data.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Culhane, D. P., Metraux, S., & Byrne, T. (2011). A prevention-centered approach to homelessness assistance: A paradigm shift? *Housing Policy Debate, 21*(2), 295–315. https://doi.org/10.1080/10511482.2010.536246

Desmond, M. (2016). Evicted: Poverty and profit in the American city. Crown Publishing.

Greenberg, G. A., & Rosenheck, R. A. (2009). Correlates of past homelessness in the National Epidemiologic Survey on Alcohol and Related Conditions. Administration and Policy in Mental Health and Mental Health Services Research, 37(4), 357–366. https://doi.org/10.1007/s10488-009-0243-x

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed.). Springer.

Kaggle. (n.d.). HUD PIT homelessness dataset. https://www.kaggle.com/datasets/bigquery/sdoh-hud-pit-homelessness

Montgomery, A. E., Fargo, J. D., Byrne, T. H., Kane, V., & Culhane, D. P. (2013). Universal screening for homelessness and risk for homelessness in the Veterans Health Administration. American Journal of Public Health, 103(S2), S210–S211. https://doi.org/10.2105/AJPH.2013.301398

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1023/A:1022643204877

U.S. Department of Housing and Urban Development. (2023). The 2023 Annual Homeless Assessment Report (AHAR) to Congress. https://www.huduser.gov/portal/sites/default/files/pdf/2023-AHAR-Part-1.pdf

U.S. Department of Housing and Urban Development. (n.d.). Point-in-Time (PIT) count methodology guide https://files.hudexchange.info/resources/documents/PIT-Count-Methodology-Guide.pdf

VanBerlo, B., Ross, M. A., Rivard, J., & Booker, R. (2021). Interpretable machine learning approaches to the prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence, 102*, 104243. https://doi.org/10.1016/j.engappai.2021.104243

Byrne, T., Montgomery, A. E., Fargo, J. D., Roberts, C. B., & Culhane, D. P. (2018). Predictive modeling of housing instability and homelessness in the Veterans Health Administration. Health Services Research, 54(1), 75–85. https://doi.org/10.1111/1475-6773.13055

Toros, H., Flaming, D., & Burns, P. (2018). Prioritizing homeless assistance using predictive algorithms: An evidence-based approach. Cityscape, 20(1), 117–135.

Heston, T. F. (2023). The cost of living index as a primary driver of homelessness in the United States. Cureus, 15(9), e45214. https://doi.org/10.7759/cureus.45214

Messier, G., Talarico, R., & Farhani, F. (2021). Predicting chronic homelessness: The importance of comparing algorithms appropriately. Journal of Social Distress and Homelessness. https://doi.org/10.1080/15228835.2021.1972502

ABOUT THE AUTHOR

Jaden Cheung is a 10th-grade student at The Athenian School in Danville, California. He is interested in data science, and machine learning, especially how predictive modeling can be used to better understand issues such as homelessness.