Homelessness in America

Regis university | Instructor John Koenig

MSDS696 Data Science Practicum II

Ashley Mimaki

**Introduction**

Previous research has shown a connection between homelessness, mental health, and drug addiction. Mental illness can make it more difficult to keep a steady income or keep up with other daily activities which in turn can make it more difficult to keep housing. For single adults’ mental illness is thought to be the third largest cause of homelessness (National Coalition for the Homeless, 2009). The inability to gain access to proper mental health care or experiencing homelessness can make it more likely for people to turn to illicit drugs which in turn can make housing more difficult to find. These are complicated issues that form a muti-way relationship (Tarr, 2018). Homelessness is a socioeconomic issue that every state in the United States faces. It can be agreed upon that homelessness is a multifaceted issue with many causes.

**Research Problem**

In this research I sought to determine if I could use a prediction model to determine a states percentage of its population that is experiencing homelessness. I needed to answer what factors affect a state’s homeless population. I also wanted to pose the question, could addressing these factors help tackle the problem of homelessness.

**Methodology**

To start I collected the data that I needed. I collected data from the Substance Abuse and Mental Health Services Administration that contained information on illicit drug use at a state level. I collected data from the US Department of Housing and Urban Development that contained information on a state’s homeless population and the count of shelters and beds available on a yearly basis within each state. I collected the total population of each state. Finally, I collected data from Mental Health America that included their ranking system of state’s mental illness prevalence as well as the state’s availability to provide treatment.

After collecting data, I observed the data sets for unnecessary points and combined them into a single usable data set for a regression model. I did this through excel as well as through Python. Some of the Excel sheets of data contained multiple years

For my exploratory data analysis, I looked at a correlation table between the factors in my data. As well as observing different graphs to compare the factors in my data.

When creating my model, as I mentioned before I used a regression model because the target of my prediction was not within a binary. First, I tested a KNN regression model, then I moved onto a decision tree model and a random forest model. These options I believe would have been better suited if I was looking for a classification model rather than the regression model. After those models I moved on to a sequential neural network model I tried multiple different activations for the model with different epoch values as well.

**Results**

When looking at the correlations between the state’s factors that I collected and the percent of a state’s population experiencing homelessness I created these visualizations:

Chart

Description automatically generated Timeline

Description automatically generated

Chart, bar chart

Description automatically generated A picture containing chart

Description automatically generated

The top left chart shows the percentage of people within an age demographic who used illicit drugs in a year organized by the percent of population experiencing homelessness in descending order. It is possible to see a slight correlation between the variables. The top right chart shows the number of beds available in shelters in the state. The bottom left shows a few factors of mental health rankings, Adult Access to Care, Mental Health Workforce Availability, and Adult Prevalance of Mental Illness, organized by the percentage of a state’s population experiencing homelessness. There is a slight correlation between the Mental Health Workforce Availability and the percentage of people experiencing homelessness. Finaly, the bottom right visual shows a correlation heatmap containing all of the variables I used to create the prediction models.

With the random forest model I used, the MSE was 0.01 and the MAE was a 0.06. Out of my KNN model and my decision tree model the random forest model had the lowest error scores. With the sequential neural network model, I used a sigmoid activation, and the model had a recorded loss value of 0.0112 and a MAE value of 0.0747. These can be seen on the graphs below. These were the lowest error values out of the different variations of neural network models that I ran.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

The MAE value means that the mean average error was about 0.07% for both models as I was predicting the percentage of the population that is experiencing homelessness. From the previous charts I presented we can see that a state’s homeless population ranges from about 0.1% to 0.9% which means the prediction model is not bad however there is room for improvement.

**Conclusion**

With my research and model, I found that it is possible to predict the percentage of a state’s population that is experiencing homelessness given certain factors. Prediction models like the one I use could be implemented when evaluating state or city projects to combat homelessness and the causes of it. This could lead to improvements in shelter availability or other factors to decrease homelessness.

Future research to aid the potential implementation of these models could include looking at these factors at a city level, this would allow for more data points which could help improve the model. Other future research could include different factors that are thought to affect homelessness to create a better model. The additional data points that I would collect for further research would be average rent prices for cities as well as average income for cities.

**Resources**

*2019-2020 National Survey on Drug Use and Health: Model-Based Prevalence Estimates (50 States and the District of Columbia) | CBHSQ Data*. (2021, December 29). Substance Abuse and Mental Health Services Administration. https://www.samhsa.gov/data/report/2019-2020-nsduh-state-prevalence-estimates

Frost, J. (2021, April 1). *Making Predictions with Regression Analysis*. Statistics By Jim. https://statisticsbyjim.com/regression/predictions-regression/

Homelessness Statistics by State, Tools for Action. (2021). *Homelessness Statistics by State | United States Interagency Council on Homelessness (USICH)*. https://www.usich.gov/tools-for-action/map/#fn[]=1300&fn[]=2900&fn[]=6400&fn[]=10200&fn[]=13400

Konkiewicz, M. (2021, December 12). *Sorting data frames in pandas - Towards Data Science*. Medium. https://towardsdatascience.com/sorting-data-frames-in-pandas-a5a3af6f346a#:%7E:text=Let%E2%80%99%20start%20with%20sorting%20a%20data%20frame%20by,using%20a%20parameter%20called%20%E2%80%98by%E2%80%99%20%3A%20df.sort\_values%20%28by%3D%27name%27%29

*Low-Income Housing Tax Credit (LIHTC): Tenant Level Data | HUD USER*. (2020). U.S. Department of Housing and Urban Development. https://www.huduser.gov/portal/datasets/lihtc/tenant.html

Mental Health America. (2019). *The State of Mental Health in America*. <https://mhanational.org/sites/default/files/2019-09/2019%20MH%20in%20America%20Final.pdf>

National Coalition for the Homeless. (2009, July). Mental Illness and Homelessness. https://www.nationalhomeless.org/factsheets/Mental\_Illness.pdf

*PIT and HIC Data Since 2007*. (2022, February). HUD Exchange. https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/

Pramoditha, R. (2021, December 16). *Train a regression model using a decision tree - Towards Data Science*. Medium. <https://towardsdatascience.com/train-a-regression-model-using-a-decision-tree-70012c22bcc1>

Tarr, P. (2020, February 3). Homelessness and Mental Illness: A Challenge to Our Society. Brain & Behavior Research Foundation. https://www.bbrfoundation.org/blog/homelessness-and-mental-illness-challenge-our-society