Tyler Benson
SML 312: Projects in Data Science
Fall 2022
Due 12/09/22

What factors best predict rates of homelessness in U.S. communities?

Overview:

Shelter is one of the most basic human necessities. In fact, the lack of it, even if the period of homelessness is short, has drastically negative impacts on the individual (Toro et al., 1991; Culhane, 2008). Many policies are made in attempts to provide support and preventative measures to homelessness in the U. S. However, homes are getting more expensive (Williams, 2021) and, even when jobs are abundant, people still are unable to find a "place to call home" (Healy, 2013). Especially when charity services are such a point of contention in politics, it is essential that homelessness services are provided in a manner that is optimal and data-driven.

The application of modern data science models allows for greater complexity in analysis of the factors that may predict homelessness. I propose that by understanding the complex relationships that rates of homelessness hold with a variety of factors through examining how machine learning models predict rates of homelessness, decisions in policy and homeless support can be made with greater confidence and effectiveness.

To accomplish this, I apply hyper parameter tuned regularized linear regression models to determine the features that are most predictive of rates of homelessness. Additionally, a modern model (a deep neural network of tree-based algorithms) called TabNet (Arik & Pfister, 2021) is applied to predict rates of homelessness and serves to explain 76% of the variance in rates of homelessness over the baseline linear regression's 54%. TabNet's feature importances are different from the linear models, and provide interesting insights as to the complex relationships that contribute to rates of homelessness. Particularly, the features with the highest predictive power were determined to be the median home value from 2016, the 2011 share of renters, the 2011 percentage of homeowners with severe cost burden, and the rate of poverty.

Related Work:

Homelessness and its history has been well studied in a variety of contexts. While homelessness has always been an issue, its prevalence in research grew in the 1980s (Grant et al., 2013). In their literature review, Grant described how homelessness has since progressed into the modern day.

Among other trends, they identified that “housing costs consumed an often unmanageable proportion of the income of poor and low-wage earning families, and have been steadily rising.” (Grant et al., 2013).

It is intuitively understandable that housing prices, in relation to family income, impacts rates of homelessness. The Statistica Research department found through a survey that "not being able to afford to buy a home was the biggest reason renters gave when asked why they didn’t currently own a home in a recent survey" (Statista Website, 2022). Government policy has attempted to respond to this with rent control law. Joe Biden has promised to ensure that “no family… would have to spend more than 30% of its income on rent” (Finnegan, 2020). Analysis into this relationship has shown that rent control laws have a small, yet relevant affect on homelessness populations (Appelbaum et al., 1991; Grimes & Chressanthis, 1997). Additionally, alternative approaches to government aid can and should be used to reduce rates of homelessness (Toro et al., 1991; Culhane, 2008). In addition to rent assistance, “supportive services coupled with permanent housing, particularly when combined with effective discharge from institutions, especially mental hospitals; (2) mediation in Housing Courts; … and (4) rapid exit from shelter” (Burt, 2005) are 4 services shown to combat homelessness effectively. Further research into the complex relationships of policy on a local level could be compelling and is a focus point of my research paper.

A primary motive for beginning this research was an anecdotal belief that rates of homelessness would be higher in locations with temperate climates. Research has affirmed this guess with rates of homelessness, specifically unsheltered homelessness, being increased in warmer climates (Appelbaum et al., 1991; Grimes & Chressanthis, 1997). Also, breaking homelessness into the subsets of warm and cold places yields interesting results, “housing prices, poverty rates, and religiosity are much more strongly associated with rates of unsheltered homelessness in warm places than in cold places” (Corinth & Lucas, 2018). Further exploration into the relationship of rates of homelessness and the climate will be another focus point through my research question.

Previous work has been done to model homelessness using a wide range of factors. Thomas Byrne et al. (2013) improved their baseline 35% explanation to 58% explanation of the variance in rates of homelessness through addition of Housing and Urban Development Point-in-Time data related to community-level homelessness. Byrne broke the data into metropolitan and non-metropolitan communities before conducting separate analyses including factors covering climate, the housing market, safety nets, and demographics. Similarly, Jamison D. Fargo et al. (2013) modeled rates of homelessness on the community-level for a single night in 2009 and accounted for 25% to 50% of the variance. To compare to these scores, my research project will employ machine learning models to explain the variance in rates of homelessness in U.S. communities using the dataset generated by Hiren Nasar et al. (2019). Nasar’s work involved extensive feature analysis that served to explain the various factors that correlate to "rates of total, sheltered, unsheltered homelessness in communities across the nation”. Their findings include conclusions that housing factors are most consistently associated with higher rates of homelessness.

In summary, these works related to rates of homelessness in the U.S. elucidate three points of focus for my research paper. Through analysis of which factors are most predictive of homelessness, we will be sure to explore the relationships between homelessness and climate, policy, and the housing market.

Relevant Data

The U.S. Department of Housing and Urban Development provides a variety of datasets related to housing and homelessness.

This research project will primarily use the dataset created by Nasar's (2019) research article. It contains 330 features of climate, housing, economic, market, demographic, and geographic domains. It will be explored in great detail in the Analysis/Modeling section. Additionally, a dataset that contains the geographic position of each Continuum of Care community will be used for graphing features onto a map of the communities in the U.S.

The San Diego Regional Library has five years of monthly geographic positions of homeless sleepers[1]. While this data may not be applicable to analyzing rates of homelessness in US communities. It allowed our analysis to extend into the individual level and served to spark the inspiration for this research paper.

Analysis/Modeling

Homeless Clustering

My inspiration for this project came from an article I read in San Diego's newspaper that stated that people were voting against homelessness services because the existence of support centers attracted more homeless people to their neighborhoods. To explore this belief that homeless people cluster and move to locations of support, I started my analysis with the San Diego Regional Library dataset of homeless sleepers. I started by generating a heatmap of downtown San Diego that showed hotspots for homeless sleepers, shown in Figure 1.

[1] https://data.sandiegodata.org/dataset/sandiegodata-org-dowtown-homeless/#packages

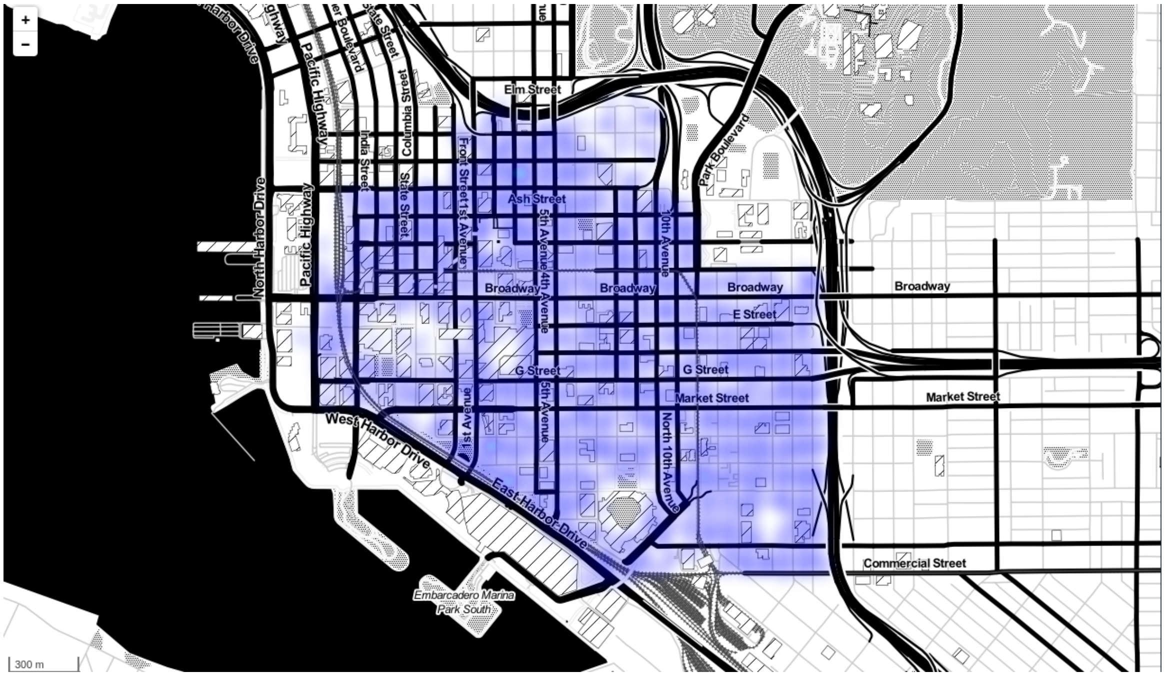


Figure 1: Heatmap of homeless sleepers from San Diego Regional Library dataset of five years of monthly homeless sleeper geographic positions.

The next step would be to gather data related to homeless services in downtown San Diego and append it to Figure 1 to compare if the hotspots on the map are clustered around the services. However, I decided that an answer to the question of "What determines the nature of homeless clustering?" isn't optimal in being able to help make decisions about large-scale policy. I liked this dataset because it represents the homeless population on an individual-level, and could allow compelling analysis into what determines how the homeless population clusters; however, I needed to broaden my scope to the community-level to better align with my goal.

Also, analysis of this kind wouldn't involve any machine learning models covered in this course. The model would involve using the statistical spread of homeless services to generate a simulated dataset of the homeless population and then performing a statistical significance test to evaluate if the simulated data was similar to the actual distribution of homeless around the provided services. The statistical significance test would be a "average nearest neighbor" test, well-described in Arc-GIS documentation[2].

[2] https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/h-how-average-nearest-neighbor-distance-spatial-st.htm

Examining the Dataset

Therefore, I transitioned my analysis to the dataset from the research article titled Market Predictors of Homelessness[3]. This dataset contains 330 features and 3009 rows of data, where each row is a different Continuum of Care (CoC) community's data for a particular year between 2010 and 2017. Because there are a substantial amount of features, I began by exploring and examining the dataset to get a better understanding. I had a few goals in mind. First, I wanted to identify a target feature, a datapoint related to homelessness that I could hold out as my dependent variable during prediction. Second, I wanted to understand the distribution of some of the predictive features so I could make data-intelligent decisions during modeling. Finally, I needed to clean the dataset to be ready for modeling.

To start, the dataset came with a data dictionary that describes each 330 features. I began by listing out the features that were labeled as "Outcome" or "Secondary Outcome" features:

Feature	Feature Definition	Domain
pit_tot_shelt_pit_hud	total sheltered - HUD PIT	Outcome
pit_tot_unshelt_pit_hud	total unsheltered - HUD PIT	Outcome
pit_tot_hless_pit_hud	total homeless - HUD PIT	Outcome
pit_miss	sum of all PIT count values	Outcome
odd_flag	odd year of data indicator	Outcome
pit_hless_balance	number of non-missing total homeless values across all years	Outcome
pit_shelt_balance	number of non-missing sheltered homeless values across all years	Outcome
pit_unshelt_balance	number of non-missing unsheltered homeless values across all years	Outcome
unbalance_flag	flag for CoCs with less than 5 years of non-missing PIT data	Outcome
pit_shelt_pit_hud_share	rate of sheltered homeless per 10,000 people	Outcome
pit_unshelt_pit_hud_share	rate of unsheltered homeless per 10,000 people	Outcome
pit_hless_pit_hud_share	rate of total homeless per 10,000 people	Outcome
missing	number of missing homeless, sheltered, and unsheltered values	Outcome
flag_d_hless	flag for missing total homeless share value in 2017 or 2013	Outcome
d_pit_hless_pit_hud_share	4-year change in pit_hless_pit_hud_share values (2017 and 2013)	Outcome
d_pit_shelt_pit_hud_share	4-year change in pit_shelt_pit_hud_share values (2017 and 2013)	Outcome
D_pit_unshelt_pit_hud_share	4-year change in pit_unshelt_pit_hud_share values (2017 and	Outcome

[3] https://www.huduser.gov/portal/publications/Market-Predictors-of-Homelessness.html

	2013)	
pit_ind_shelt_pit_hud	individuals sheltered - HUD PIT	Secondary Outcome
pit_ind_unshelt_pit_hud	individuals unsheltered - HUD PIT	Secondary Outcome
pit_ind_hless_pit_hud	total individuals - HUD PIT	Secondary Outcome

Figure 2: Table of the descriptions of each “Outcome”feature in our dataset. This table only represents a subset of the outcome features. There are 36 Secondary and Primary Outcome features in total, see the full table in Appendix 1.

It will be important to remember these columns and make sure that they are not present in our modeling. Also, Figure 2 has an outcome feature titled “missing” that defines the number of missing values for reported counts of homelessness. Noticing this, I knew that I needed to understand and deal with missing values present in our dataset.

Missing Values

In particular, there are some leftover features from the research paper that generated this dataset that will not be relevant to our modeling. To begin, the reference paper performed feature generation to analyze the changes in community-level factors between 2010 and 2017. Because these features could only be defined for data collected during the year 2017, this creates a lot of missing values. I wanted to get a better understanding of the quantity and distribution of missing values, so I plotted a missing value graph that represents NaN values as transparent bars and present data as black bars as Figure 3.

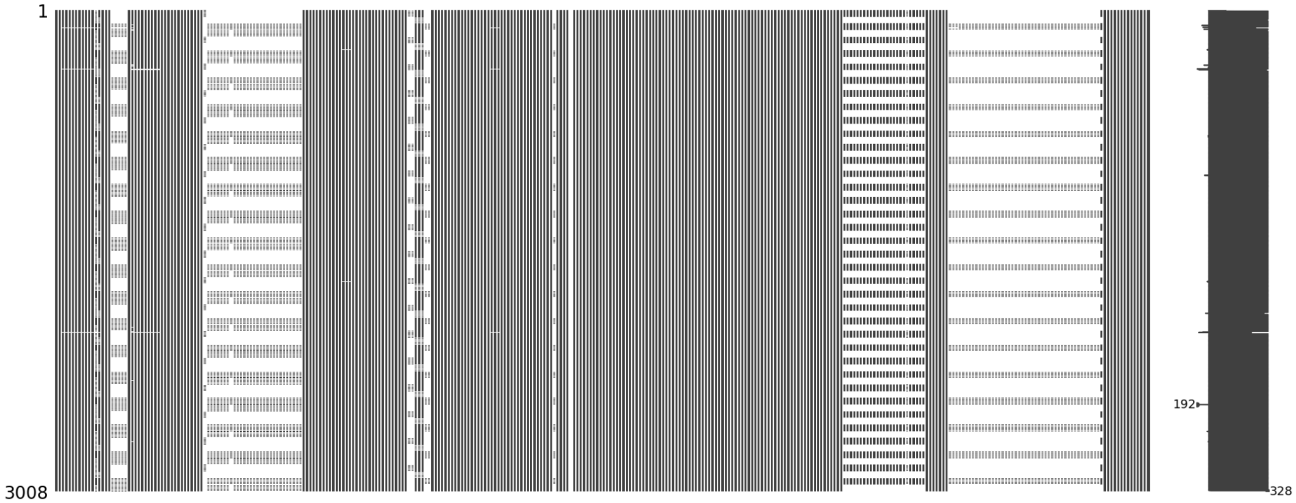


Figure 3: Missing values present in our dataset. This figure shows each row (3009 in total) and each column (330 in total) of the dataset, where each position is blank if there is a missing value and black if not.

Supporting my theory that there were a lot of leftover features that represent changes between years, Figure 3 showed me that there is a pattern to missing values. Additionally, there are some features that simply have no real values. Ultimately there are 136 columns with missing values in them. To see which columns have the most missing values, I output the column names sorted by percentage of missing values as Figure 4.

Rank	Feature	Fraction of Missing Values
1	dem_health_ins_acs5yr_2012	1
2	d_pit_hless_pit_hud_share	0.8759973404
3	d_pit_shelt_pit_hud_share	0.8759973404
4	d_pit_unshelt_pit_hud_share	0.8759973404
5	d_hou_pol_occhudunit_psh_hud	0.875
6	d_dem_pop_child_census_share	0.875
7	d_econ_labor_unemp_rate_BLS	0.875
55	hou_pol_numret12mos_hud	0.7509973404
56	hou_pol_totalind_hud	0.7509973404
57	hou_pol_numret6mos_hud	0.7509973404
58	hou_pol_totalexit_hud	0.7509973404
59	hou_pol_totalday_hud	0.7509973404
60	dem_health_ins_xt	0.75
61	dem_soc_vet_acs5yr	0.75
62	hou_mkt_utility_acs5yr	0.75
63	total_rent_inventory_acs5yr	0.75

Figure 4: Table of the fraction of missing valuesfor the top most sparsely populated features in the dataset. “Rank” represents the i’th most sparsely populated feature. The fraction of missing values was calculated by dividing the number of columns with NaNs by 3008, the number of rows.

Our inspection of the top missing value columns revealed that some of these features represent the yearly changes in other factors. These features will have lots of missing values by nature of their generation. These features are not relevant for our analysis and so I dropped the 46 columns. This left me with 90 columns to fix; because the dataset is robust and our inspection of the features implied that these columns are often feature generations in response to specific questions that the research paper determined, I decided to drop these columns as well. However, some of these columns are outcome features that need to remain in our dataset. For these features I decided to only drop the rows in which an outcome feature was NaN. After handling missing values, the dataset was reduced from 3008 by 330 to 2994 by 189.

Now that we have handled missing values, it's necessary to understand the remaining 194 features that are in this dataset. Because there are so many, simply memorizing the feature definitions from the data dictionary won't be sufficient. To start, we can output the distribution of the domains of the features in our dataset. Additionally, see Appendix 2 for the definitions of each of the 189 features.

Domain	**Counts of Features**
Demographic	55
Housing	51
Economic	23
Safety Net	16
Subgroup	14
Outcome	14
Local Policy	5
Identifier	4
Geography	4
Climate	3

Figure 5: Table of counts of the different types offeatures after missing data was handled.

In Figure 5, we can see that there is still some cleaning that needs to be done before we are ready for modeling. That is, no identifier feature should be allowed to have predictive power because we don't want the model to simply memorize which communities have a history of homelessness. Additionally, we need to pick which "Outcome" feature we want to predict. Inspecting the outcome variable definitions in Figure 2, `pit_hless_pit_hud_share` stood out to me. This variable represents the number of homeless per 10,000 population and therefore is normalized. This normalization will allow our model to avoid biases towards cities with higher populations (and thus a disproportionately higher homeless population).

Now that I selected my target variable, I wanted to familiarize myself as best as possible with its relationship with our dataset. To begin, I plotted a line graph of how the target variable changed over time for 4 personally relevant communities. With such a large dataset it was necessary that I hold some of the variables constant when graphing to augment interpretability. To briefly explain the personal relevance of these communities: NJ-514 is the community in which Princeton resides, IN-502 is where I grew up in Indiana, CA-601 is where I now live in San Diego, and HI-501 is where my girlfriend lives in Honolulu (and was one of the communities that inspired this research project).

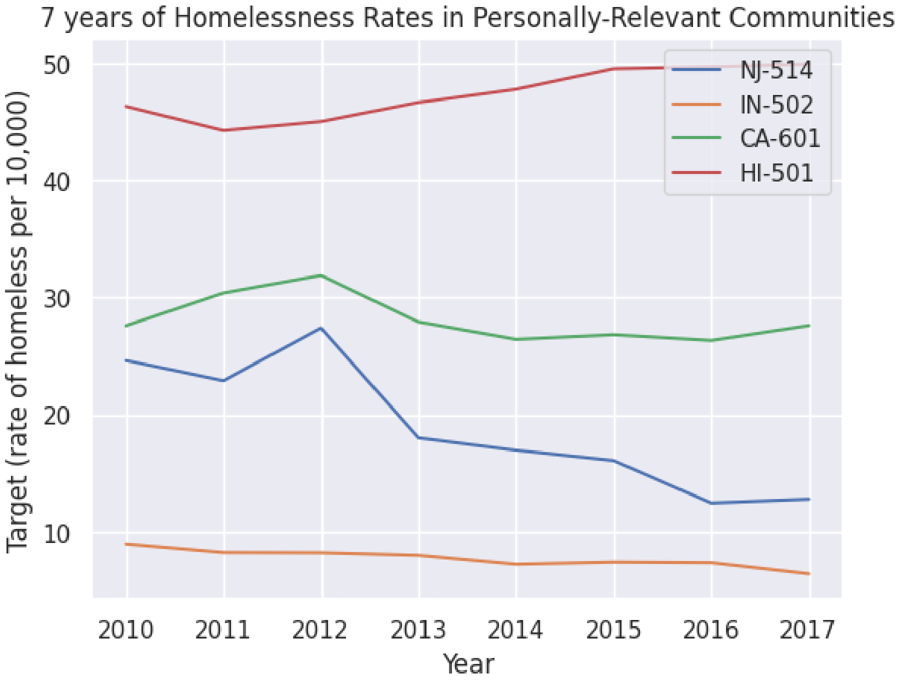


Figure 6: A line graph showing the progression of rates of homelessness for 4 personally-relevant communities between the years 2010 and 2017.

Figure 6 shows us how the rates of homelessness stay relatively constant year to year, with some fluctuation from an average for each community. Also, we can see that of these 4 communities, the community in Honolulu, Hawaii has the highest rate of homelessness.

Looking at Appendix 2 of feature definitions, our dataset includes some indicator variables that could help segment the data for visualization. It could be interesting to compare distributions of our target variable across these indicator variables. To do this, I created 4 violin plots that each had a violin for both outcomes of the indicator variable 0 and 1.

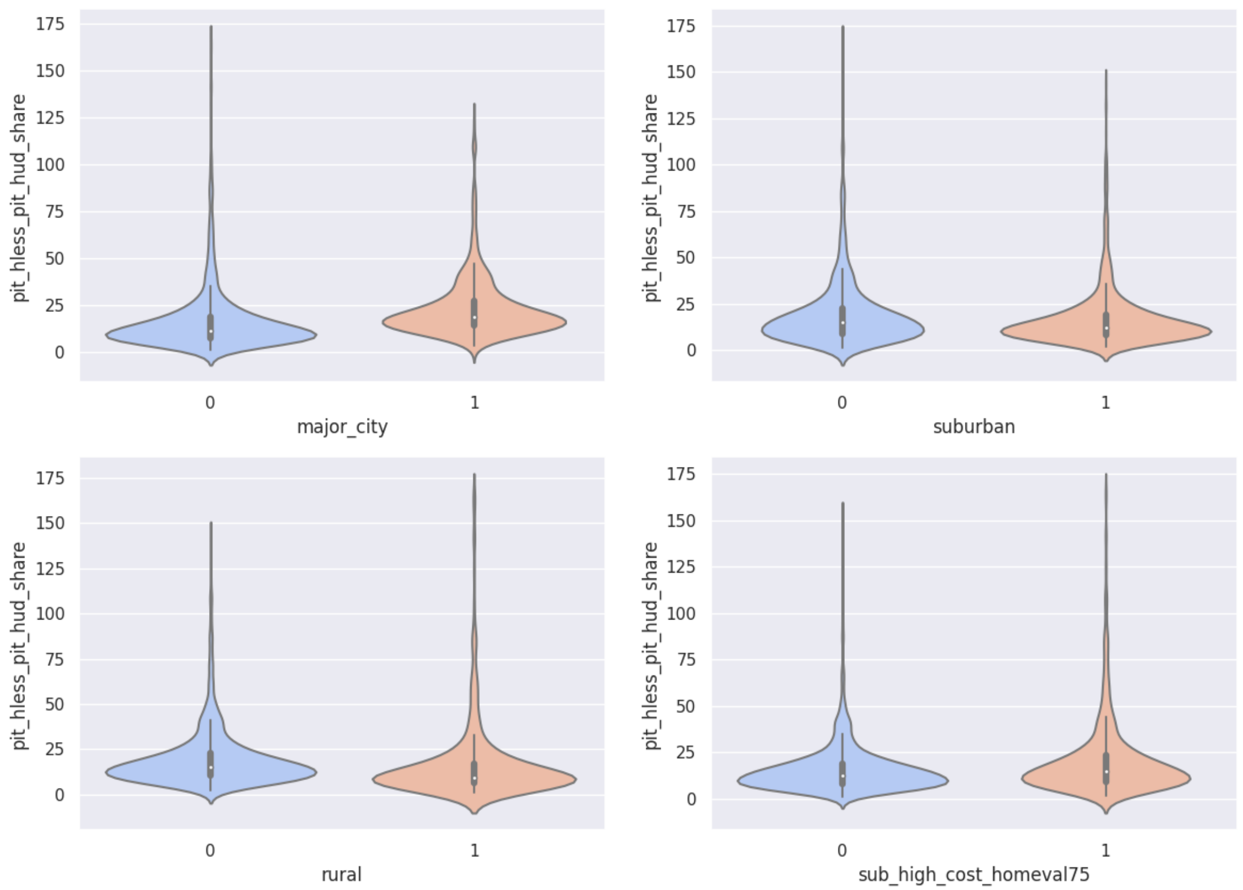


Figure 7: Violin plots for the indicator variables major_city, suburban, rural, and ub_high_cost_homeval75. For each of the indicator variables, the distribution of the target variable is shown as a violin.

Figure 7 implies that the distribution of the target variable may be slightly different depending on some of our indicator variables. For example, the two features of rural and major cities seem to shift the distribution of the target (the mode of the distribution is slightly different).

Outliers

These graphs show that there may exist some outliers in our target dataset. Let's take a look at these outliers -- defined to have a target value more than 2 standard deviations away from the mean. In this case, we are only looking at the outliers on the upper side of rates of homelessness, in which the rates of homelessness are significantly higher than average. I initially included the lower outlier rates of homelessness; however, our target set is heavily skewed and with a long right tail. With a mean of 18.75 and a standard deviation of 18.45, no communities are an outlier for low rates of homelessness.

The cutoff for communities that represent outliers in terms of rates of homelessness was selected to be 3 times the standard deviation above the mean. After segmenting the data to only consist of these outlier communities, I performed additional exploratory analyses to see if there was any noticeable difference in features for these outlier communities. To begin, I listed out which community was most likely to have an outlier rate of homelessness.

Continuum of Care Number (identifier)	Counts
CA-509	8
FL-604	8
DC-500	8
FL-518	7
CA-508	6
CA-501	5
NY-600	5
CA-614	5
CA-504	4
FL-519	3
FL-505	3
CA-522	3
FL-517	2
LA-503	2
CA-603	1
CA-524	1
MD-508	1
MO-602	1
NY-607	1

Figure 8: Table showing the counts of outlier communities in our dataset. Because there were 8 years of data collected, the highest count possible is 8.

Figure 8 showed the likelihood of a given community to be an outlier community. Not all of the communities were outliers every single year, yet some were; these communities to keep in mind are `CA-509`, `DC-500`, `FL-604`. Taking a look at all of the outlier communities, there seemed to be some states that are more likely to be outliers, let's take a look at the distribution of states of the outlier rows.

State	Counts
CA	33

FL	23
DC	8
NY	6
LA	2
MD	1
MO	1

Figure 9: Table showing the counts of the states that the outlier communities reside in.

As expected, there were only a few states that are often extreme outliers for rates of homelessness. Taking a look at the states with the highest amount of extreme outliers, California and Florida have a few similarities that supported the initial hypothesis for this research paper. For example, both states have relatively warm weather. This may suggest that including average temperature + temperature fluctuation year over year for each community may improve predictability.

Geospatial Graphs

It is important to be able to visualize our CoC communities to see exactly what each row of data in our dataset means. To do this I imported the CoC regional map dataset to map the data geospatially. This geographic dataset had a lot of similarities to the dataset that we have been working with so far; however, the data was collected in 2021. I'd never worked with geospatial data, so, to get a better understanding of how to graph data geographically I made a graph that plotted one of the features in the geospatial dataset that seemed relevant to our analysis in Figure 10 below.

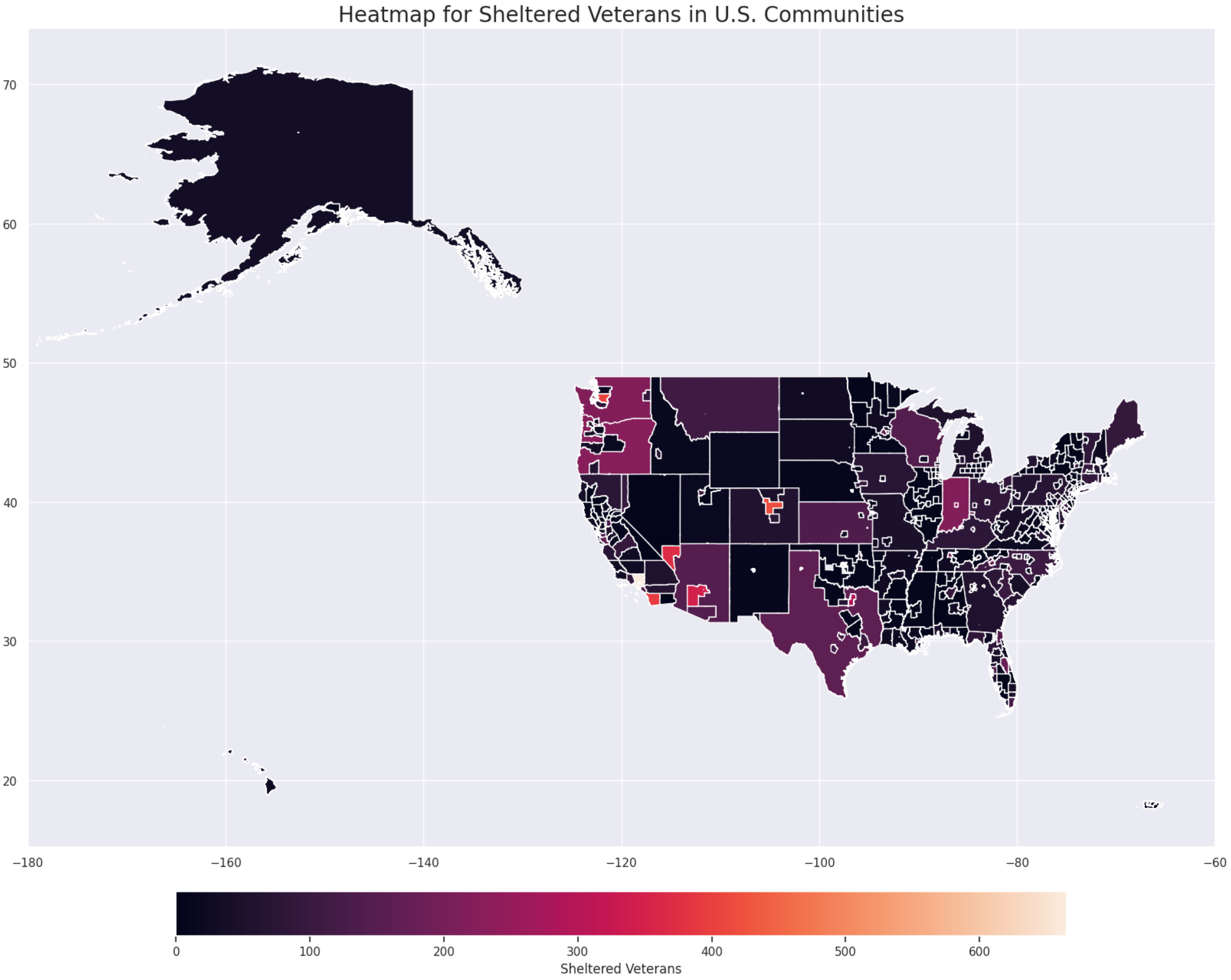


Figure 10: A geographical heat-map of the number of sheltered veterans in each U.S. Community from 2021.

Continuing the plan to understand the outlier communities, I created a map of the US and highlighted the top 10 communities that were most likely to be outlier communities in terms of rates of homelessness per 10,000 people in Figure 11 below. Consistent with my analysis of outlier communities before, we can see that the outlier communities are predominantly present in the states California and Florida. Additionally, this map allows us to notice that almost all of the outlier communities are coastal. It may be interesting to engineer an indicator variable for whether or not the community is a coastal community to aid in our prediction.

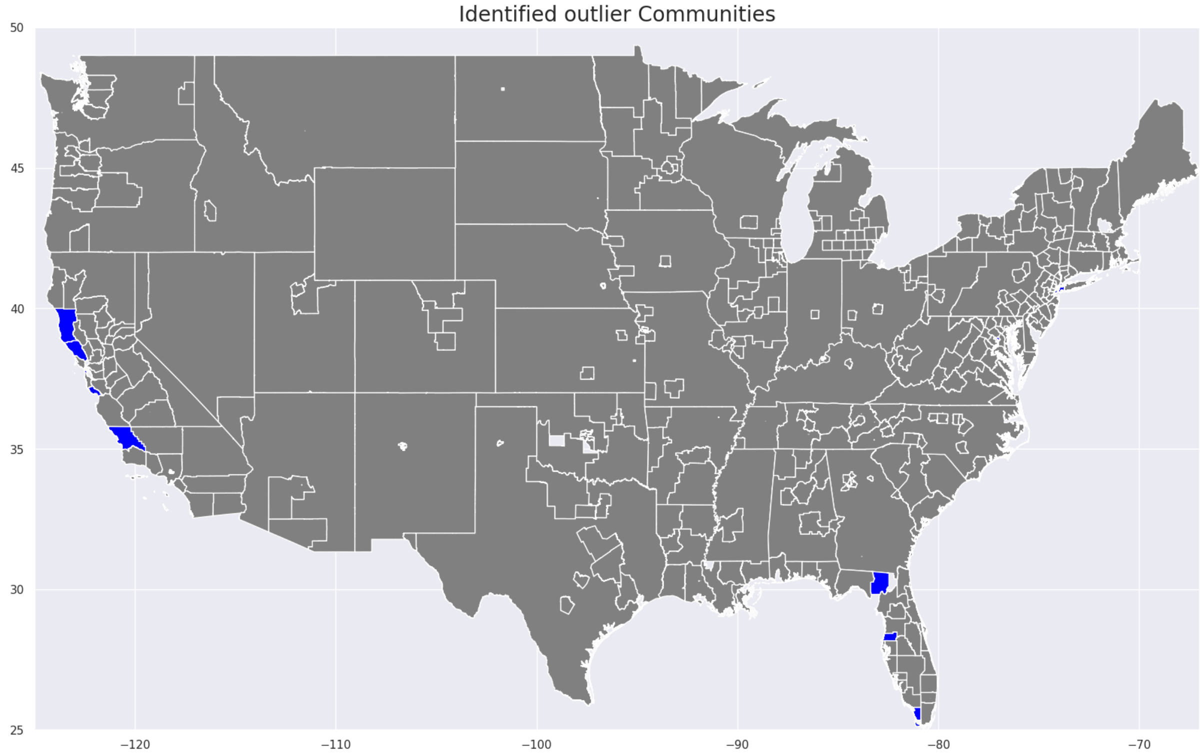


Figure 11: A map of the U.S. with identified outlier communities highlighted in blue. Hawaii, Alaska, and Puerto Rico aren't shown as they are not outliers.

Correlation

With 100+ predictor columns, I was in need of performing feature analysis. Some questions I wanted to answer were: which features are the most correlated with the target? Which non-target features are cross-correlated and might impact a multi-variable linear model?

I started by plotting a correlation heatmap matrix of our columns appended as Figure 12 Subplot (a). Inspecting this heatmap it can be seen that there are some independent variables that are highly cross-correlated. For example, the top left corner of this map, appended as Figure 12 Subplot (b), includes demographic data that is obviously cross-correlated. It is to be expected that a feature defining the female population is almost perfectly correlated with a feature for male population census data.

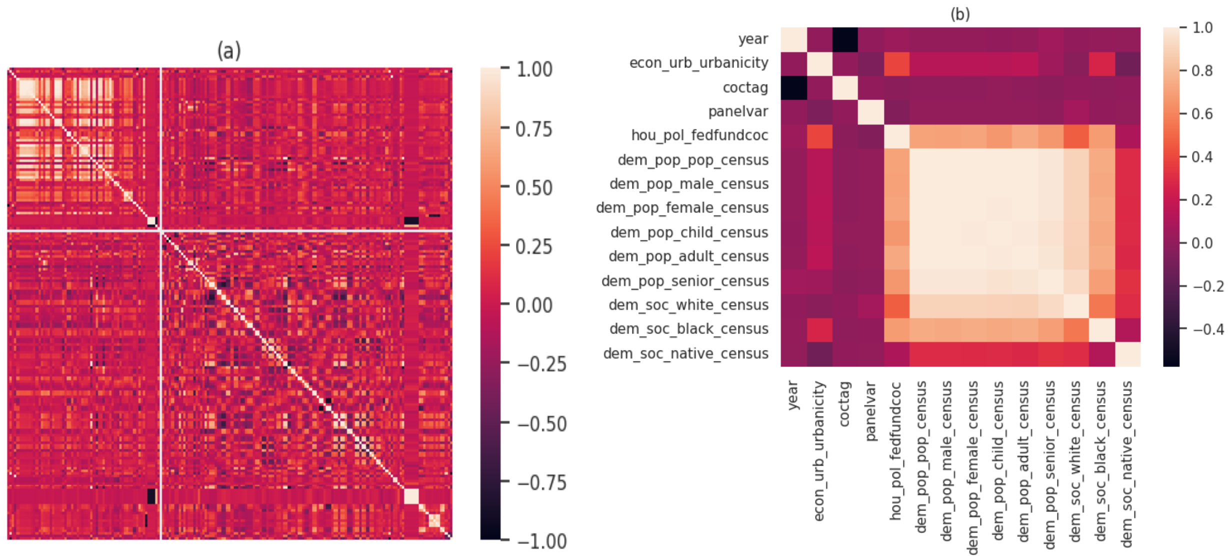


Figure 12: Subplot (a) shows the correlation heatmap for all the numeric variables in our feature-set. Subplot (b) shows the top left corner of Subplot (a), using only the first 15 features in correlation analysis.

It is pretty difficult to determine the specific correlations with our target feature from this correlation map. So I plotted the top 20 features that have the highest correlation in absolute value with our target feature. In our final model, however, we won't have access to "Outcome" variables that are in this data. So before I returned these top features I dropped outcome features that we weren't going to be using in prediction.

Feature	Correlation with Target (absolute)
hou_mkt_rentshare_acs5yr_2012	0.3988829353
hou_mkt_burden_sev_own_acs_2012	0.3953996081
hou_mkt_rentshare_acs5yr_2017	0.3918018901
hou_mkt_burden_sev_own_acs_2017	0.3563331941
sub_west_coast	0.3544097623
dem_health_alcdeath_IMHE_2015	0.3408802225
hou_mkt_homeval_acs5yr_2012	0.3391770465
hou_mkt_burden_sev_own_acs_diff	0.3362158882
sub_west_census	0.3271389176
hou_mkt_homeval_acs5yr_2017	0.3265020779
hou_mkt_ovrcrowd_acs5yr_2017	0.3163065895
econ_labor_topskill_acs5yr_2012	0.3141764234
env_wea_avgtemp_noaa	0.2974310528
econ_labor_incineq_acs5yr_2017	0.2895409995

hou_mkt_ovrcrowd_acs5yr_2012	0.2886156793
hou_mkt_burden_sev_rent_acs_2017	0.2863372412
econ_labor_midskill_acs5yr_2017	0.2855501698
census_division	0.2811489682
census_region	0.277348604

Figure 13: The top 20 features in absolute value of their correlation with the target feature of ‘pit_hless_pit_hud_share’.

Similarly, I looked at the distribution of types of features that are most highly correlated with the target. This will help us determine what domains are often correlated with rates of homelessness.

Domain	**Counts in Top Correlated Features**
Housing	10
Economic	3
Geography	2
Subgroup	2
Climate	1
Demographic	1

Figure 14: The distribution of the domains that our highly correlated features reside in.

We can make a couple observations from this. First, it is significant that of the top 20 correlated features, 10 of them are related to the housing domain. Additionally, of the 3 climate features present in our dataset, 1 of them is in the top 20. To speak honestly, I expected more climate features to be highly correlated with rates of homelessness; however, one is better than none. Also, none of the features in the “Local Policy” domain are present in the top 20. This may imply that the relationship between local policy and rates of homelessness is either more complex than a simple correlation can reveal, or non-existent.

We established before through Figure 12 that there is some obvious collinearity within our feature-set. Multicollinearity could reduce the explainability of our model as well as could introduce overfitting problems for some machine learning models, including linear regressions. To combat this, we could perform feature selection. That is, we could reduce redundancy in our dataset by taking only one of any two features that are collinear. In fact, I started to run an analysis of the variance inflation factor on our dataset as a means to test for multicollinearity; however this took an extremely long time on our expansive set of features and may not be necessary. That is, as long as we avoid making conclusions based on the coefficients of a regression model in our analysis and are careful about our conclusions from the explainability in multi-variable linear regression models, our analysis will be sound. Additionally, as shown later in the Modeling subsection, principal component analysis is performed as a means to reduce collinearity, and yields worse performance. This implies that the models are robust to the collinearity present in our dataset.

Before we were ready for modeling, some final data cleaning was necessary. I needed to remove the identifier features that were leftover from joining tables. These included 4 columns to remove; the most relevant identifiers to our analysis were the year and the identifiers that signify which particular community the row of data is collected from. Additionally, we needed to inspect and deal with categorical variables. In our dataset, there is only one categorical feature of state. Fortunately, this dataset already includes indicators for regions and in this way has already encoded state and so the decision was made to drop the one categorical variable.

Modeling

The data was split into a hold-out test set (20%) and a training set (80%) to allow for final model evaluations. For each iteration in the modeling stage, the same schema for cross validation was used for model selection: A five (5) fold cross validation was performed on the training set and the average and standard deviation of the training and testing scores were returned. The metric used in scoring was selected to be R-squared because it was common among the related works to score modeling based on how much the variance in homelessness is explained by the feature-set. The average R-squared across these cross validation folds for each of the models is shown in Figure 15 for each model.

Model	**Avg R2 Train**	**Std R2 Train**	**Avg R2 Val**	**Std R2 Val**
Linear Regression	0.644	0.01318	0.571	0.0700
Ridge Regression	0.625	0.01352	0.562	0.0645
Ridge Regression Hyperparameter Tuned	0.653	0.01350	0.581	0.0653
Lasso Regression	0.537	0.01189	0.502	0.05263
Lasso Regression Hyperparameter Tuned	0.6286	0.01344	0.5649	0.0646
TabNet	0.9241	0.0472	0.8285	0.032

Figure 15: The results from cross-validation for the six iterations of modeling. The average (Avg) and standard deviation (Std) R-squared (R2) metrics for the five-folds of cross-validation are shown for both the train (Train) and validation (Val) sets. The models are ordered top to bottom in terms of chronological implementation.

All of the following modeling, except for TabNet, was done using scikit-learn's packages (Pedregosa et al., 2011). First, I started by making a baseline model of Linear Regression. This linear model will help to compare my analysis with related works and serve to prove that a mistake wasn't made during data cleaning. As shown in Figure 15, the linear regression model obtained an average R-squared score of 0.571 on the validation set and a 0.644 on the train set during cross validation.

The test score was lower than the train score by a relevant margin, suggesting that the linear model is overfitting during training. To follow this theory, I applied a regularized linear regression model called Ridge Regression to the dataset to see if a regularizer can reduce overfitting and improve our test score.

It can be seen in Figure 15 that Ridge regression did reduce training score as expected but the testing score was not improved over the baseline model. It could be possible that the model needs a more fine-tuned amount of regularization. Ridge uses the parameter 'alpha' in order to determine the amount of regularization applied during model training – higher values of alpha lead to harsher regularization. In order to determine the best amount of regularization used in the Ridge model, I applied GridSearchCV[4] that performs cross-validation with alphas ranging from 0-1000. Figure 15 shows the scores for this optimized model, obtaining better scores than the baseline in both training and validation sets.

However, the GridSearchCV output the optimized parameters for Ridge regression to be alpha to equal 0. This was unexpected. When alpha is 0, the Ridge regression model is equivalent to a linear regression model[5], that is, the model no longer regularizes during training. First off, it is immediately apparent and confusing that the Ridge regression performed better on average than the linear regressor did, even when alpha is 0. This is most likely due to the other parameter that was tuned during hyperparameter tuning of normalization=True; the model must perform better when it normalizes the incoming data.

Alternatively concerning why the model may perform better in the absence of ridge regularization: ridge regression intrinsically penalizes for having coefficients close to zero, and so an alternative theory was that our dataset isn't optimized for making every coefficient non-zero. To test this theory, I decided to apply a regularizer that optimizes for fewer non-zero coefficients, known as Lasso regression. If Lasso regression hyperparameter tunes to a nonzero alpha, I would be more confident in a reasoning that our dataset is optimized for using the most amount of features during prediction.

Unfortunately, however, Lasso regression also hyper-parameter-tuned alpha to be 0. Additionally, as can be seen in Figure 15, Lasso regression performed worse than the baseline both before and after hyperparameter tuning. With this in mind, we may be able to reason that our dataset is not optimized for using every single feature during prediction and rather performs better when the best features are identified and their predictive power emphasized.

A key issue with this dataset is its cross-collinearity between the independent variables. One such approach to resolving this issue is through principal component analysis (PCA). PCA reduces the dimensionality and redundancy of the feature space through singular value decomposition. However, the contribution of individual factors to each of the reduced components is difficult to verify and

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

thus feature importances output from a model using the PCA dataset will be essentially uninterpretable. Despite this, it could prove interesting to see if model performance is improved through solving the multi-collinearity problem. To start, the data was reduced from 171 independent variables to 130 principal components. The parameter of 130 components was selected because it reduced multicollinearity while maintaining relevant R-squared scores. To show how the PCA dataset no longer is collinear, a heatmap was generated.

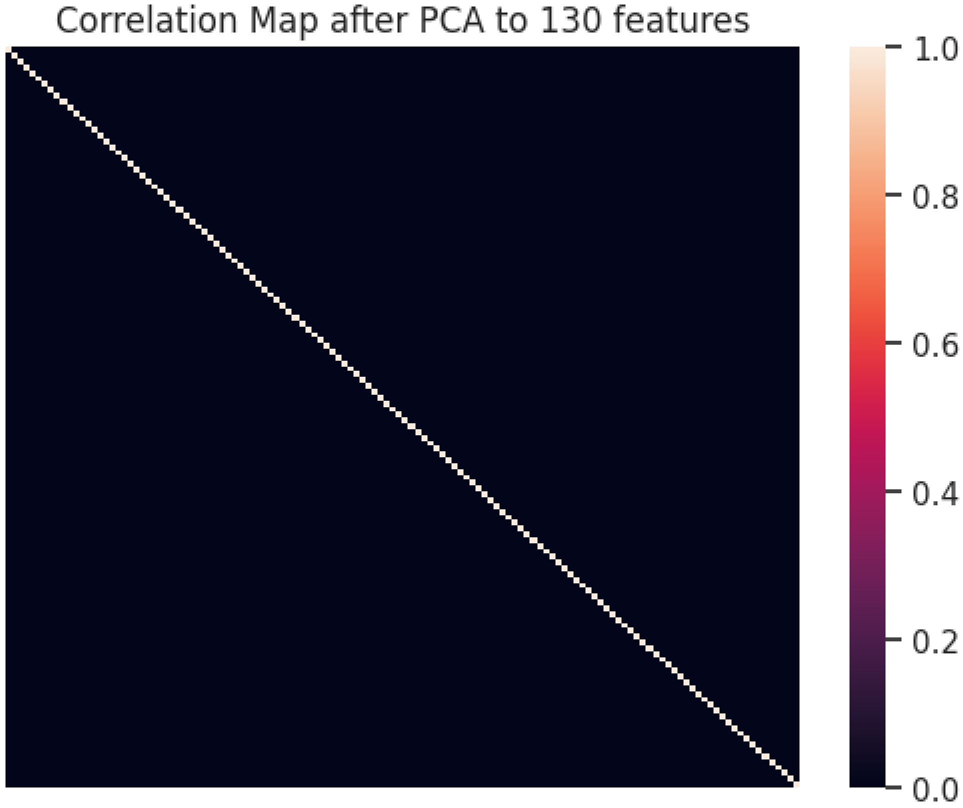


Figure 16: Correlation heatmap for the PCA dataset.

It is obvious that the components no longer correlate with each other from this Figure 16. To see if reducing dimensionality and collinearity was effective, the cross-validation steps were performed for Linear and Ridge regressions. The results of this are shown in Figure 17 below. Unfortunately, the results of each of the models were worse than the models before PCA was applied. Because of this, the decision was made that collinearity is not drastically affecting model performance and that PCA will not be applied.

Model	**Avg R2 Train**	**Std R2 Train**	**Avg R2 Val**	**Std R2 Val**
Linear Regression	0.623	0.0124	0.545	0.0607
Ridge Regression	0.623	0.0124	0.547	0.0596

Figure 17: Results for five fold cross-validation for the models on the PCA dataset.

I was worried that the interpretability of the feature importances output from any of the linear models previously implemented would be influenced by the cross-correlation present in our dataset. That is, I wanted to apply a model that would allow for greater complexity in relationships analyzed in our dataset as well as greater interpretability. This initially made me think of tree-based algorithms, whose interpretability is as simple as following the "tree" to see why the model made the decision it did. Looking for a model that matched all of these characteristics brought me to TabNet[6]. TabNet is a model developed by Google in 2019 that employs deep neural networks in combination with tree-based algorithms to allow for high-interpretability and efficient training[7]. As seen in Figure 15, TabNet scored considerably better than the baseline model (and the hyperparameter tuned Ridge regression) on both the train and validation set on average. Interestingly, if we compare the standard deviations of the distribution of train and test scores, we can identify that TabNet is performing more *reliably* than the linear models. TabNet achieved a 0.03 standard deviation in R-squared scores on the validation set, while every other model implemented achieved above a 0.06.

Finally, I evaluated the baseline model, the hyperparameter tuned Ridge regression model, and TabNet on the initial hold-out set after training it on the entire train set to allow for final model comparison. This is shown in Figure 18 below.

Model	**R2 Train**	**R2 Test**
Linear Regression	0.640	0.541
Ridge Regression Hyperparameter Tuned	0.623	0.538
TabNet	0.911	0.766

Figure 18: Final model evaluations on both the train and hold-out test sets after being trained on the entire train set.

Interestingly, the hyperparameter tuned Ridge regression performed worse on the test set than the baseline model did, even though it performed better on the validation set on average during cross-validation. This is most likely due to random chance as the difference is only 0.003.

Results

The research question fundamental to this paper involves analyzing the complex relationships that factors may have with rates of homelessness on a community level. In order to respond to this, I output the feature importances for both the hyperparameter-tuned Ridge Regression model and the TabNet model. These two models were chosen because they scored well either during final evaluation or during cross-validation – a higher scoring will allow greater confidence in the important features that the models output.

[6] https://ojs.aaai.org/index.php/AAAI/article/view/16826

[7] https://www.geeksforgeeks.org/tabnet/

Feature Importances from Ridge Regression

To start, I performed feature importance analysis on the hyperparameter-tuned Ridge Regression model. Unfortunately Ridge Regression doesn't have an internal method for returning feature importances. While coefficients are present and could be analyzed in comparison to each other, Ridge penalizes large coefficients, therefore influencing the interpreted importances in prediction. Additionally, coefficients are only interpretable if the incoming data is normalized, which was not done during data cleaning. Instead, permutation importance was used to determine which feature holds the most predictive power. Permutation importance fits and evaluates the model on permutations of subsets of features to determine which features are most important to model performance. The results are shown in Figure 19 below.

Feature	Feature Definition	Mean Importance
dem_pop_male_census	total male population, intercensal estimate	12.418
dem_pop_female_census	total female population, intercensal estimate	7.960
dem_soc_white_census	total white alone (non-hispanic) population, intercensal estimate	5.021
dem_pop_adult_census	total population ages 20-64, intercensal estimate	3.858
econ_labor_force_pop_BLS	total population in the labor force	2.11
econ_labor_emp_pop_BLS	total employed population	2.026
hou_mkt_homeage1940_acs5yr_2017	2016 percentage of housing units built before 1940	2.003
hou_mkt_units_census	total number of housing units, intercensal estimate	1.489
dem_pop_pop_census	total population, intercensal estimate	1.260
dem_soc_hispanic_census	total latino/hispanic (all races) population, intercensal estimate	1.173
hou_mkt_homeval_acs5yr_2012	2011 median home value	0.906
hou_mkt_homeage1940_acs5yr_2012	2011 percentage of housing units built before 1940	0.811
dem_soc_black_census	total black alone (non-hispanic) population, intercensal estimate	0.538
econ_labor_medinc_acs5yr_2012	2011 median income	0.449
hou_mkt_homeage_acs5yr_2012	2011 median housing unit age	0.363
hou_mkt_medrent_acs5yr_2017	2016 median rent	0.356
dem_pop_child_census	total population ages 0-19, intercensal estimate	0.337
econ_sn_ssdi_SSA	total number of disabled workers receiving benefits	0.297
dem_mort_lifeexp_IMHE_2015	2014 life expectancy, from IMHE	0.220
hou_mkt_renter_count_evlab	count of renter-occupied households, eviction lab estimate from census and ESRI	0.220

Figure 19: The top 20 features as determined by permutation importance and their respective mean permutation importance for the hyperparameter tuned Ridge Regression model. Feature definitions are included for ease of reference.

It is additionally insightful to look at the distribution of top features in our results. This is shown in Figure 20 below.

Feature Domain	Counts
Demographic	9
Housing	5
Economic	3
Safety Net	3

Figure 20: The distribution of the domain of the top20 features determined by permutation importance.

As informative as it is to see the top 20 most predictive features for Ridge regression in predicting rates of homelessness, I decided it may prove interesting to inspect the bottom 20 as well as their domains to include in the results. These are shown in Figures 21 and 22 below.

Feature	Feature Definition
econ_labor_unemp_pop_BLS	total unemployed population
dem_health_mhlth_chr_share_2017	2016 share of mental health care providers to total population
hou_mkt_density_census	housing units estimate divided by square miles
fhfa_hpi_2009	2009 base year house price index (HPI), from FHFA
evict_flag	flag for missing eviction rate value
sub_tight_high_cost_rent	indicator for tight, high cost rental market CoCs
dem_soc_singadult_acs5yr_2012	2011 percentage of children living with a single parent
hou_pol_hlessconduct_total	total count of prohibited conduct laws
dem_pop_mig_census	net migration from year-1 to year, intercensal estimate
econ_sn_cashasst_acs5yr_diff	change in econ_sn_cashasst_acs5yr values (2016 and 2011)
sub_low_rent_vacancy	indicator for CoCs with rental vacancy rates <= 5 percent
hou_mkt_ovrcrowd_acs5yr_diff	change in hou_mkt_ovrcrowd_acs5yr values (2016 and 2011)
env_wea_precip_noaa	total January precipitation
dem_pop_mig_census_share	yearly increase in population to total population
dem_soc_ed_bach_acs5yr_2012	2011 share of the population with bachelors or higher
hou_mkt_density_dummy	indicator for high housing density CoC (>= 75th percentile)
hou_mkt_rentvacancy_acs5yr_2012	2011 rental vacancy rate
fhfa_hpi_flag	flag indicating that counties in the CoC had missing variables for housing price

hou_mkt_rentvacancy_acs5yr_2017	2016 rental vacancy rate
econ_labor_incineq_acs5yr_diff	change in econ_labor_incineq_acs5yr values (2016 and 2011)
econ_labor_unskill_acs5yr_diff	change in econ_labor_unskilled_acs5yr values (2016 and 2011)

Figure 21: The bottom 20 features as determined by permutation importance of the hyperparameter tuned Ridge Regression model. These are ordered top to bottom in terms of least important. Feature definitions are included for ease of reference.

Feature Domain	**Counts**
Housing	8
Demographic	5
Economic	3
Subgroup	2
Local Policy	1
Climate	1
Safety Net	1

Figure 22: The distribution of the domain of the bottom20 features determined by permutation importance performed on a hyperparameter tuned Ridge Regression model.

One of the core hypotheses was that features that involve weather will hold predictive power for rates of homelessness. Contrarily, none of the top 20 features in Figure 20 determined through permutation importance for the Ridge Regression model include weather-based statistics. Looking at the bottom 20 features in Figure 22, the weather-based feature of "total January precipitation" is interesting as to its unpredictability.

Additionally, looking at Figure 19's definitions, there seem to be three cohorts for important features. Those that describe census data and data regarding the population, features for describing the work-force and unemployment rates, and features for representing the housing market. Each of these cohorts make sense as to why they would be predictive of rates of homelessness. Particularly, the feature of median home value is well documented to be correlated with homelessness. However, of the 20 features that were highest in absolute value correlation with our target variable only `hou_mkt_homeval_acs5yr_2012` was determined to be an important feature in Ridge Regression.

Another core component of this research paper involved policy and its relationship with rates of homelessness. `hou_pol_hlessconduct_total` as the count of prohibited conduct law violations is present in the 20 most unimportant features for Ridge Regression. The Australian Human Rights Commision cites Jones v City of Los Angeles in saying "Courts in the United States have held that these types of laws violate the constitutional right to freedom from cruel and unusual punishment because they punish homeless people on the basis of their status, not because of their conduct[8]. `hou_pol_hlessconduct_total` represents counts of laws that could be defined as such, vague definitions of prohibited conduct that allows Police to enact anti-homelessness policies. In theory I would expect this feature to be predictive of homelessness. Either communities with a high count of anti-homeless laws are enacting those laws in response to high rates of homelessness, or the laws work and homelessness decreases… either way a complex relationship would seem to exist.

However, permutation importance as a means for determining feature importance suffers from cross-correlated feature spaces (Vorotyntsev, 2020). To start, some of the top 20 features identified through permutation importance are cross-correlated with each other. Take for example the two features of `dem_pop_male_census` and `dem_pop_female_census`, if the male population increases, it can be supposed that the female population also increases – it is obvious how these two features are highly correlated. To support this, the correlation heat map from Figure 12 shows that a high amount of our independent variables are cross-correlated.

While this doesn't completely invalidate the analysis we've performed, it would be interesting to see if applying a different model that contains an intrinsic method for determining feature importances and is known for its explainability outputs differing important features. This supported my decision to model feature importances with TabNet for comparison and insight into the complex relationships surrounding rates of homelessness.

Feature Importances from TabNet

TabNet's internal feature importances method was used to output the following Figure 23.

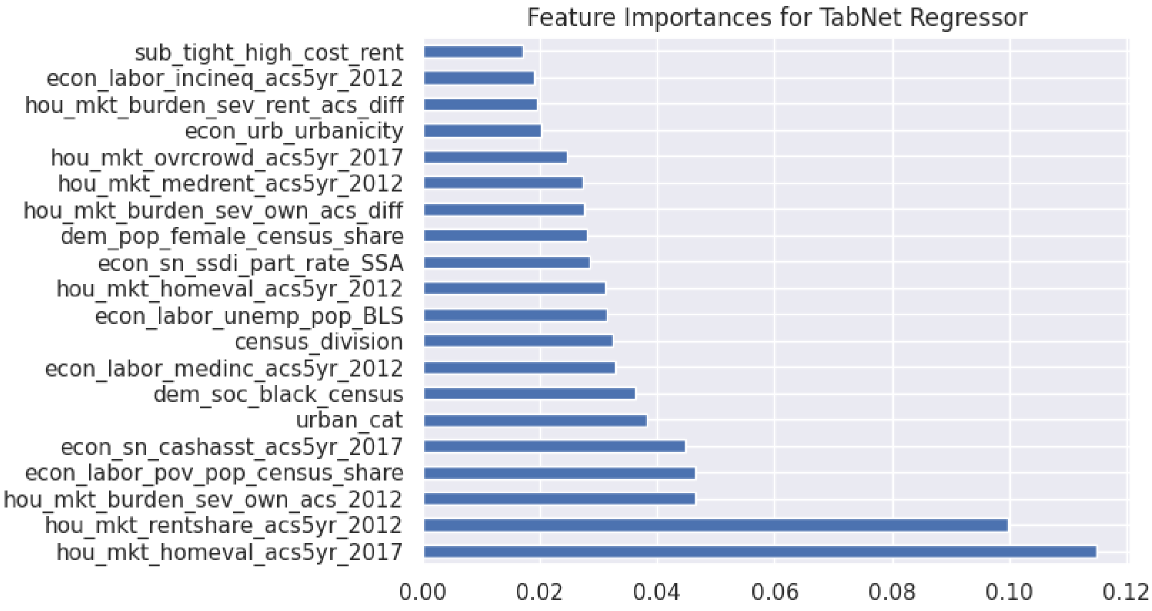


Figure 23: Feature importances for the TabNet regression model trained on the entire training set.

[8] https://humanrights.gov.au/our-work/rights-and-freedoms/publications/homelessness-human-rights-issue

Feature	Feature Definition	Importance
hou_mkt_homeval_acs5yr_2017	2016 median home value	0.1150
hou_mkt_rentshare_acs5yr_2012	2011 share of renters	0.0998
hou_mkt_burden_sev_own_acs_2012	2011 percentage of home owners with severe cost burden	0.0466
econ_labor_pov_pop_census_share	poverty rate, number of persons in poverty to total population	0.0465
econ_sn_cashasst_acs5yr_2017	2016 share of households with public assistance income	0.0448
urban_cat	urbanicity category	0.0383
dem_soc_black_census	total black alone (non-hispanic) population, intercensal estimate	0.0364
econ_labor_medinc_acs5yr_2012	2011 median income	0.0329
census_division	census region	0.0323
econ_labor_unemp_pop_BLS	total unemployed population	0.0314
hou_mkt_homeval_acs5yr_2012	2011 median home value	0.0311
econ_sn_ssdi_part_rate_SSA	SSDI participation rate (ssdi participants/census population)	0.0285
dem_pop_female_census_share	female population to total population	0.0280
hou_mkt_burden_sev_own_acs_diff	change in hou_mkt_burden_sev_own_acs5yr values (2016 and 2011)	0.0275
hou_mkt_medrent_acs5yr_2012	2011 median rent	0.0272
hou_mkt_ovrcrowd_acs5yr_2017	2016 share of overcrowded housing units	0.0245
econ_urb_urbanicity	urbanicity category	0.0201
hou_mkt_burden_sev_rent_acs_diff	change in hou_mkt_burden_sev_rent_acs5yr values (2016 and 2011)	0.0195
econ_labor_incineq_acs5yr_2012	2011 gini coefficient	0.0191
sub_tight_high_cost_rent	indicator for tight, high cost rental market CoCs	0.0171

Figure 24: The top 20 features as determined by feature importance of TabNet ordered top to bottom in terms of least important. Feature definitions are included for ease of reference.

Feature Domain	Counts
Housing	8
Economic	4
Subgroup	3
Demographic	2
Safety net	2

Geography	1

Figure 25: The distribution of the domain of thetop 20 features output by feature importance of the TabNet model.

Excitingly, TabNet reduced the amount of demographic-type features from 10 to 2 in the most important feature set. As explained before, some of these demographic features are obviously highly-correlated, and thus were incorrectly deemed important through permutation importance. This reduction allows greater confidence in the feature importances output by TabNet.

Three of the top features include the rental-housing market as highly predictive for rates of homelessness. This reflects the research explored in the Related Works section. It is interesting to see that the rental-housing market is as predictive as it is expected to be correlated with rates of homelessness. In fact, seven (7) of TabNet's top 20 predictive features are present in the top 20 features in absolute value correlation with the target (TabNet's top 3 are all in the top 20 of the features with high correlation). Because TabNet scored significantly higher than the baseline models, the result that the top 3 predictive features were simply highly correlated with the target may suggest that TabNet is able to understand some deeper relationships that lie below the surface. As it stands, TabNet wasn't developed to print out the decision trees that contribute to its decision making, so future research that employs such a method may garner greater insight into these complex relationships.

None of the top 20 features were in the "Local Policy" domain. Considering that one of this research paper's goals was to aid in data-driven policy, this is a little disappointing; however, this suggests some interesting areas for future research. Performing the same modeling and determining feature importances on subsets of the data that depend on variations in local policy could begin to explore the relationship that policy has on rates of homelessness.

Interestingly, 11 of the 20 top features are from single years. For example, grouping the dataset by CoC identifier, it is apparent that each community has the same 2016 median home value for each year of its recorded data. However, our dataset didn't include median home values for each particular year. Therefore, a reasoning that median home values are significantly predictive of rates of homelessness even in the presence of large amounts of redundancy arises. Future research that is able to include each year's median home value in modeling rates of homelessness may help justify this claim.

To further understand the top features output by TabNet, additional exploratory data analysis was performed. Comparing the target feature with the top predictors directly may show interesting results. To begin, the target feature was plotted as a heatmap using the geographic shapes from the geographic dataset as Figure 26.

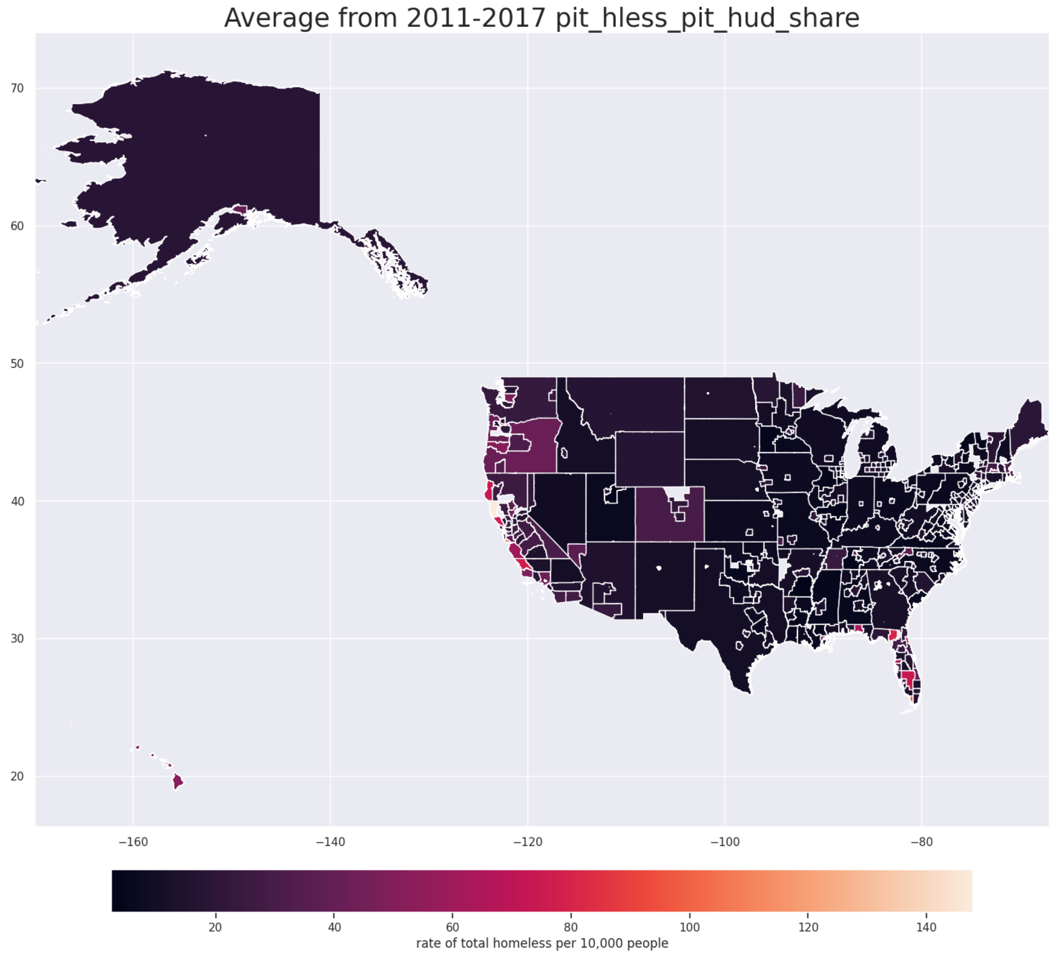

Average from 2011-2017 pit_hless_pit_hud_share
70
60
50
40
30
20
−160
−140
−120
−100
−80
20
40
60
80
100
120
140
rate of total homeless per 10,000 people

Figure 26: A heatmap of rates of total homeless per10,000 people for each community in the U.S. The data for the target feature is averaged for each year.

To compare, the same plot was made using the average median home value from 2016, the top predictor determined by Tabnet, shown in Figure 27 below, as well as for the next 3 most predictive features shown in Appendix 3, 4, and 5.

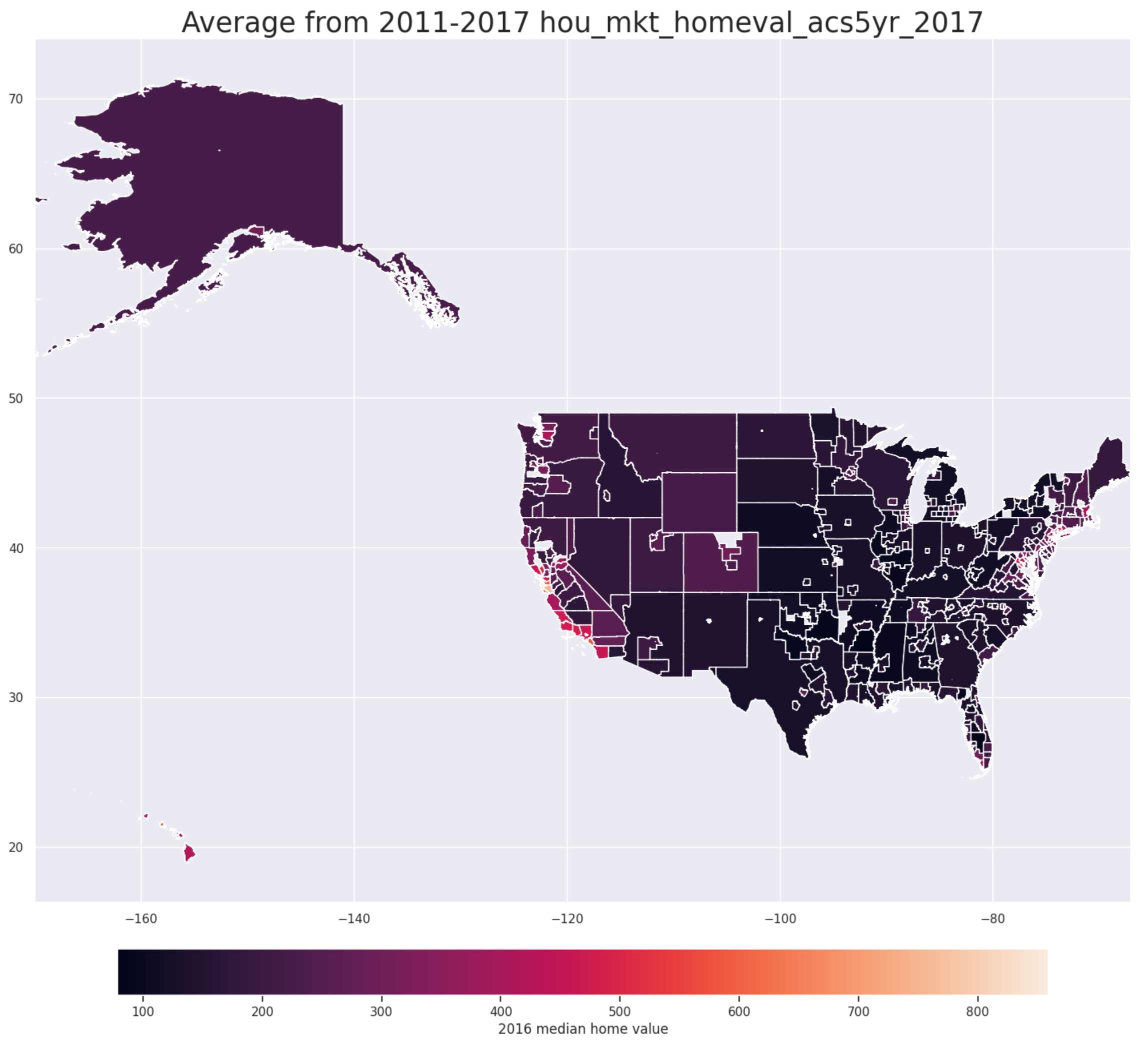


Figure 27: A heatmap of rates of total homeless per10,000 people for each community in the U.S. The data for the target feature is averaged for each year.

From visual inspection, it can be seen that the median home value from 2016 holds similarities to rates of homelessness. Specifically, looking at the west coast of California, these two Figures 26 and 27 show almost the same pattern of spikes in value along these communities. Appendix 3 holds similarities in the Florida communities and could account for the outlier communities that are present in that state.

Conclusions

The factors most predictive for rates of homelessness in U.S. communities were determined to be the domains of the housing market and the economy. My focus was on factors related to climate, local policy, and the housing market and, in conclusion, climate and local policy factors were not as predictive as expected for rates of homelessness, while the housing market was most predictive. Additionally, TabNet considerably improved the baseline model performances and showed better performance than models in related works. TabNet's application and interpretability could prove interesting in future works on the analysis of rates of homelessness.

Appendix

Feature	Feature Definition	Domain
pit_tot_shelt_pit_hud	total sheltered - HUD PIT	Outcome
pit_tot_unshelt_pit_hud	total unsheltered - HUD PIT	Outcome
pit_tot_hless_pit_hud	total homeless - HUD PIT	Outcome
pit_miss	sum of all PIT count values	Outcome
odd_flag	odd year of data indicator	Outcome
pit_hless_balance	number of non-missing total homeless values across all years	Outcome
pit_shelt_balance	number of non-missing sheltered homeless values across all years	Outcome
pit_unshelt_balance	number of non-missing unsheltered homeless values across all years	Outcome
unbalance_flag	flag for CoCs with less than 5 years of non-missing PIT data	Outcome
pit_shelt_pit_hud_share	rate of sheltered homeless per 10,000 people	Outcome
pit_unshelt_pit_hud_share	rate of unsheltered homeless per 10,000 people	Outcome
pit_hless_pit_hud_share	rate of total homeless per 10,000 people	Outcome
missing	number of missing homeless, sheltered, and unsheltered values	Outcome
flag_d_hless	flag for missing total homeless share value in 2017 or 2013	Outcome
flag_xt_hless	flag for missing total homeless share value in an odd year	Outcome
flag_d_shelt	flag for missing sheltered homeless share in 2017 or 2013	Outcome
flag_xt_shelt	flag for missing sheltered homeless share value in an odd year	Outcome
flag_d_unshelt	flag for missing unsheltered homeless share in 2017 or 2013	Outcome
flag_xt_unshelt	flag for missing unsheltered homeless share vale in an odd year	Outcome
d_pit_hless_pit_hud_share	4-year change in pit_hless_pit_hud_share values (2017 and 2013)	Outcome
d_pit_shelt_pit_hud_share	4-year change in pit_shelt_pit_hud_share values (2017 and 2013)	Outcome
D_pit_unshelt_pit_hud_share	4-year change in pit_unshelt_pit_hud_share values (2017 and 2013)	Outcome
pit_ind_shelt_pit_hud	individuals sheltered - HUD PIT	Secondary Outcome

pit_ind_unshelt_pit_hud	individuals unsheltered - HUD PIT	Secondary Outcome
pit_ind_hless_pit_hud	total individuals - HUD PIT	Secondary Outcome
pit_perfam_shelt_pit_hud	persons in families sheltered - HUD PIT	Secondary Outcome
pit_perfam_unshelt_pit_hud	persons in Families unsheltered - HUD PIT	Secondary Outcome
pit_perfam_hless_pit_hud	total persons in families - HUD PIT	Secondary Outcome
pit_ind_chronic_hless_pit_hud	total chronically homeless individuals - HUD PIT	Secondary Outcome
pit_perfam_chronic_hless_pit_hud	total chronically homeless persons in families - HUD PIT	Secondary Outcome
pit_vet_hless_pit_hud	total veterans - HUD PIT	Secondary Outcome
hou_pol_totalind_hud	length of time homeless; ES, SH, and TH universe	Secondary Outcome
hou_pol_totalday_hud	length of time homeless; ES, SH, and TH total days	Secondary Outcome
hou_pol_totalexit_hud	total exits: universe	Secondary Outcome
hou_pol_numret6mos_hud	total exits: return in less than 6 months	Secondary Outcome
hou_pol_numret12mos_hud	total exits: return in less than 12 months	Secondary Outcome

Appendix 1: Table of the feature definitions of theoutcome features in our dataset.

Feature	**Feature Definition**	**Associated Domain**
year	year	Identifier
cocnumber	continuum of care number	Identifier
econ_urb_urbanicity	urbanicity category	Subgroup
coctag	tag(cocnumber)	Identifier
panelvar	CoC number used to set panel	Identifier
hou_pol_fedfundcoc	CoC federal funding - HUD	Safety Net
dem_pop_pop_census	total population, intercensal estimate	Demographic
dem_pop_male_census	total male population, intercensal estimate	Demographic

dem_pop_female_census	total female population, intercensal estimate	Demographic
dem_pop_child_census	total population ages 0-19, intercensal estimate	Demographic
dem_pop_adult_census	total population ages 20-64, intercensal estimate	Demographic
dem_pop_senior_census	total population ages 65 or older, intercensal estimate	Demographic
dem_soc_white_census	total white alone (non-hispanic) population, intercensal estimate	Demographic
dem_soc_black_census	total black alone (non-hispanic) population, intercensal estimate	Demographic
dem_soc_native_census	total native alone (non-hispanic) population, intercensal estimate	Demographic
dem_soc_asian_census	total asian alone (non-hispanic) population, intercensal estimate	Demographic
dem_soc_pacific_census	total pacific islander alone (non-hispanic) population, intercensal estimate	Demographic
dem_soc_racetwo_census	total population of two or more races (non-hispanic), intercensal estimate	Demographic
dem_soc_hispanic_census	total latino/hispanic (all races) population, intercensal estimate	Demographic
fhfa_hpi_flag	flag indicating that counties in the CoC had missing variables for housing price	Housing
fhfa_hpi_2009	2009 base year house price index (HPI), from FHFA	Housing
econ_labor_force_pop_BLS	total population in the labor force	Economic
econ_labor_emp_pop_BLS	total employed population	Economic
econ_labor_unemp_pop_BLS	total unemployed population	Economic
econ_labor_unemp_rate_BLS	unemployment rate, from BLS LAUS	Economic
hou_mkt_area_census	area in square miles - land area	Housing
dem_pop_density_census	total population estimate divided by square miles	Demographic
hou_mkt_density_census	housing units estimate divided by square miles	Housing
dem_pop_mig_census	net migration from year-1 to year, intercensal estimate	Demographic
hou_pol_hudunit_psh_hud	number of HUD-assisted units	Safety Net
hou_pol_occhudunit_psh_hud	HUD unit occupancy rate	Safety Net
hou_mkt_units_census	total number of housing units, intercensal estimate	Housing
hou_mkt_pmt_totbldg_census	total permitted buildings	Housing
hou_mkt_pmt_totunit_census	total permitted units	Housing
hou_mkt_pmt_totvalue_census	total value of permitted buildings	Housing
econ_labor_pov_pop_census	count of all ages in poverty, from SAIPE	Economic
econ_labor_medinc_census	median household income, from SAIPE	Economic

econ_sn_ssdi_SSA	total number of disabled workers receiving benefits	Safety Net
econ_sn_ssdi_part_rate_SSA	SSDI participation rate (ssdi participants/census population)	Safety Net
econ_sn_ssi_part_SSA	total number of SSI recipients	Safety Net
econ_sn_ssi_pay_SSA	amount of payments in thousands of dollars	Safety Net
econ_sn_ssi_part_rate_SSA	SSI participation rate (ssi participants/census population)	Safety Net
hou_mkt_evict_count	number of addresses with an eviction judgment, from eviction lab	Housing
hou_mkt_evict_file_count	eviction cases filed, including multiple against same address, from eviction lab	Housing
hou_mkt_renter_count_evlab	count of renter-occupied households, eviction lab estimate from census and ESRI	Housing
hou_mkt_evict_flag	flag for eviction estimates lower than the actual number, from eviction lab	Housing
hou_mkt_evict_file_rate	share of renter-occupied households with an eviction filed, from eviction lab	Housing
hou_mkt_evict_rate	share of renter-occupied households with an eviction, from eviction lab	Housing
dem_health_cost_dart	total medicare reimbursements per enrollee (parts a&b), dartmouth	Demographic
hou_pol_hlessconduct_food	count of food sharing laws	Local Policy
hou_pol_hlessconduct_total	total count of prohibited conduct laws	Local Policy
hou_pol_hlessconduct_sleep	count of sleeping, camping, lying/sitting, and vehicle restriction laws	Local Policy
hou_pol_hlessconduct_loiter	count of loitering and vagrancy laws	Local Policy
hou_pol_hlessconduct_beg	count of begging laws	Local Policy
env_wea_precip_annual_noaa	total annual precipitation	Climate
cpi_2017	consumer price index	Economic
env_wea_avgtemp_noaa	average January temperature	Climate
env_wea_precip_noaa	total January precipitation	Climate
state_abr	state abbreviation	Geography
pit_miss	sum of all PIT count values	Outcome
odd_flag	odd year of data indicator	Outcome
pit_hless_balance	number of non-missing total homeless values across all years	Outcome
pit_shelt_balance	number of non-missing sheltered homeless values across all years	Outcome
pit_unshelt_balance	number of non-missing unsheltered homeless values	Outcome

	across all years	
unbalance_flag	flag for CoCs with less than 5 years of non-missing PIT data	Outcome
pit_hless_pit_hud_share	rate of total homeless per 10,000 people	Outcome
missing	number of missing homeless, sheltered, and unsheltered values	Outcome
dem_pop_adult_census_share	adult population (ages 20-64) to total population	Demographic
dem_pop_child_census_share	child population (ages 0-19) to total population	Demographic
dem_pop_female_census_share	female population to total population	Demographic
dem_pop_male_census_share	male population to total population	Demographic
dem_pop_mig_census_share	yearly increase in population to total population	Demographic
dem_pop_senior_census_share	senior population (ages 65 and older) to total population	Demographic
dem_age_boom_census_2011	2010 share of baby boomers to total population	Demographic
dem_soc_asian_census_share	asian (non-hispanic) population to total population	Demographic
dem_soc_black_census_share	black (non-hispanic) population to total population	Demographic
dem_soc_hispanic_census_share	hispanic/latino population (all races) to total population	Demographic
dem_soc_native_census_share	native (non-hispanic) population to total population	Demographic
dem_soc_pacific_census_share	pacific islander (non-hispanic) population to total population	Demographic
dem_soc_racetwo_census_share	two or more races to total population	Demographic
dem_soc_white_census_share	white (non-hispanic) population to total population	Demographic
dem_soc_other_census_share	sum of native, pacific islander, and two or more race shares	Demographic
dem_health_mhlth_chr_share_2017	2016 share of mental health care providers to total population	Demographic
dem_health_excesdrink_chr_2017	2016 share of excess drinkers	Demographic
dem_health_alcdeath_IMHE_2015	2014 alcohol morality rate	Demographic
dem_mort_lifeexp_IMHE_2015	2014 life expectancy, from IMHE	Demographic
dem_health_ins_acs5yr_2017	2016 share of the population with health insurance	Demographic
dem_soc_ed_bach_acs5yr_2012	2011 share of the population with bachelors or higher	Demographic
dem_soc_ed_bach_acs5yr_2017	2016 share of the population with bachelors or higher	Demographic
dem_soc_ed_bach_acs5yr_diff	change in dem_soc_ed_bach_acs5yr values (2016 and 2011)	Demographic
dem_soc_ed_somecoll_acs5yr_2012	2011 share of the population with some college	Demographic
dem_soc_ed_somecoll_acs5yr_2017	2016 share of the population with some college	Demographic

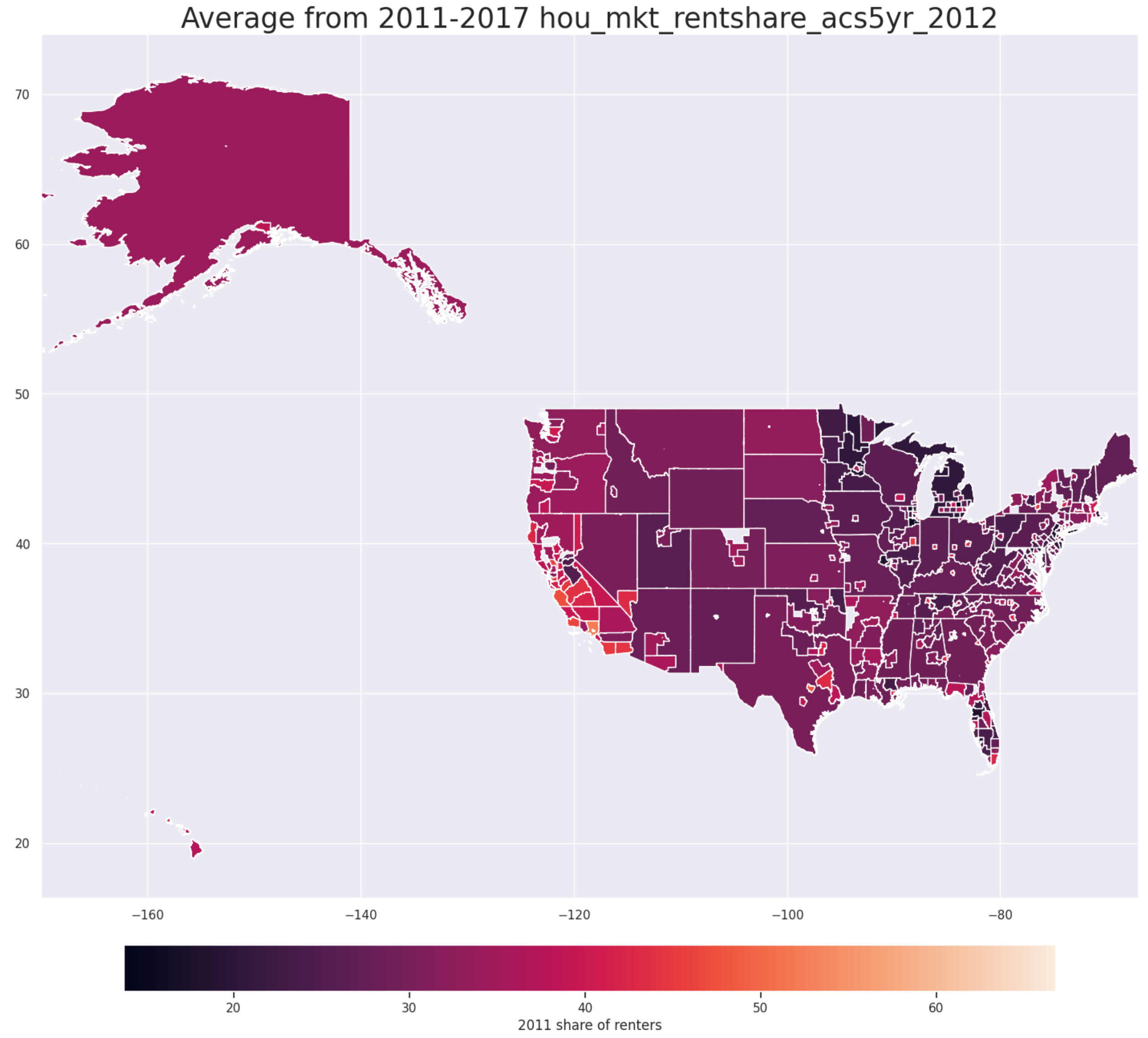
dem_soc_ed_hsgrad_acs5yr_2012	2011 high school graduate share of the population	Demographic
dem_soc_ed_hsgrad_acs5yr_2017	2016 high school graduate share of the population	Demographic
dem_soc_ed_lesshs_acs5yr_2012	2011 share of the population without high school diploma	Demographic
dem_soc_ed_lesshs_acs5yr_2017	2016 share of the population without high school diploma	Demographic
dem_soc_ed_lesshs_acs5yr_diff	change in dem_soc_ed_lesshs_acs5yr values (2016 and 2011)	Demographic
dem_soc_singparent_acs5yr_2012	2011 percentage of children living with a single parent	Demographic
dem_soc_singparent_acs5yr_2017	2016 percentage of children living with a single parent	Demographic
dem_soc_singparent_acs5yr_diff	change in dem_soc_singparent_acs5yr values (2016 and 2011)	Demographic
dem_soc_singadult_acs5yr_2012	2011 percentage of children living with a single parent	Demographic
dem_soc_singadult_acs5yr_2017	2016 percentage of children living with a single parent	Demographic
dem_soc_singadult_acs5yr_diff	change in dem_soc_singadult_acs5yr values (2016 and 2011)	Demographic
dem_soc_vet_acs5yr_2012	2011 percentage of veterans	Demographic
dem_soc_vet_acs5yr_2017	2016 percentage of veterans	Demographic
dem_soc_vet_acs5yr_diff	change in dem_soc_vet_acs5yr values (2016 and 2011)	Demographic
econ_labor_incineq_acs5yr_2012	2011 gini coefficient	Economic
econ_labor_incineq_acs5yr_2017	2016 gini coefficient	Economic
econ_labor_incineq_acs5yr_diff	change in econ_labor_incineq_acs5yr values (2016 and 2011)	Economic
econ_labor_topskill_acs5yr_2012	2011 employment rate for high-skilled workers	Economic
econ_labor_topskill_acs5yr_2017	2016 employment rate for high-skilled workers	Economic
econ_labor_topskill_acs5yr_diff	change in econ_labor_topskilled_acs5yr values (2016 and 2011)	Economic
econ_labor_midskill_acs5yr_2012	2011 employment rate for middle-skilled workers	Economic
econ_labor_midskill_acs5yr_2017	2016 employment rate for middle-skilled workers	Economic
econ_labor_midskill_acs5yr_diff	change in econ_labor_midskilled_acs5yr values (2016 and 2011)	Economic
econ_labor_unskill_acs5yr_2012	2011 employment rate for low-skilled workers	Economic
econ_labor_unskill_acs5yr_2017	2016 employment rate for low-skilled workers	Economic
econ_labor_unskill_acs5yr_diff	change in econ_labor_unskilled_acs5yr values (2016 and 2011)	Economic
econ_labor_medinc_acs5yr_2012	2011 median income	Economic
econ_labor_medinc_acs5yr_2017	2016 median income	Economic
econ_labor_pov_pop_census_share	poverty rate, number of persons in poverty to total	Economic

	population	
econ_sn_cashasst_acs5yr_2012	2011 share of households with public assistance income	Safety Net
econ_sn_cashasst_acs5yr_2017	2016 share of households with public assistance income	Safety Net
econ_sn_cashasst_acs5yr_diff	change in econ_sn_cashasst_acs5yr values (2016 and 2011)	Safety Net
hou_mkt_rentshare_acs5yr_2012	2011 share of renters	Housing
hou_mkt_rentshare_acs5yr_2017	2016 share of renters	Housing
hou_mkt_rentshare_acs5yr_diff	change in hou_mkt_rentshare_acs5yr values (2016 and 2011)	Housing
hou_mkt_rentvacancy_acs5yr_2012	2011 rental vacancy rate	Housing
hou_mkt_rentvacancy_acs5yr_2017	2016 rental vacancy rate	Housing
hou_mkt_rentvacancy_acs5yr_diff	change in hou_mkt_rentvacancy_acs5yr values (2016 and 2011)	Housing
hou_mkt_ovrcrowd_acs5yr_2012	2011 share of overcrowded housing units	Housing
hou_mkt_ovrcrowd_acs5yr_2017	2016 share of overcrowded housing units	Housing
hou_mkt_ovrcrowd_acs5yr_diff	change in hou_mkt_ovrcrowd_acs5yr values (2016 and 2011)	Housing
hou_mkt_homeage_acs5yr_2012	2011 median housing unit age	Housing
hou_mkt_homeage_acs5yr_2017	2016 median housing unit age	Housing
hou_mkt_homeage_acs5yr_diff	change in hou_mkt_homeage_acs5yr values (2016 and 2011)	Housing
hou_mkt_homeage1940_acs5yr_2012	2011 percentage of housing units built before 1940	Safety Net
hou_mkt_homeage1940_acs5yr_2017	2016 percentage of housing units built before 1940	Safety Net
hou_mkt_homeage1940_acs5yr_diff	change in hou_mkt_homeage1940_acs5yr values (2016 and 2011)	Safety Net
hou_mkt_homeval_acs5yr_2012	2011 median home value	Housing
hou_mkt_homeval_acs5yr_2017	2016 median home value	Housing
hou_mkt_homeval_acs5yr_diff	change in hou_mkt_homeval_acs5yr values (2016 and 2011)	Housing
hou_mkt_medrent_acs5yr_2012	2011 median rent	Housing
hou_mkt_medrent_acs5yr_2017	2016 median rent	Housing
hou_mkt_medrent_acs5yr_diff	change in hou_mkt_medrent_acs5yr values (2016 and 2011)	Housing
hou_mkt_utility_acs5yr_2012	2011 utility costs	Housing
hou_mkt_utility_acs5yr_2017	2016 utility costs	Housing
hou_mkt_utility_acs5yr_diff	change in hou_mkt_utility_acs5yr values (2016 and	Housing

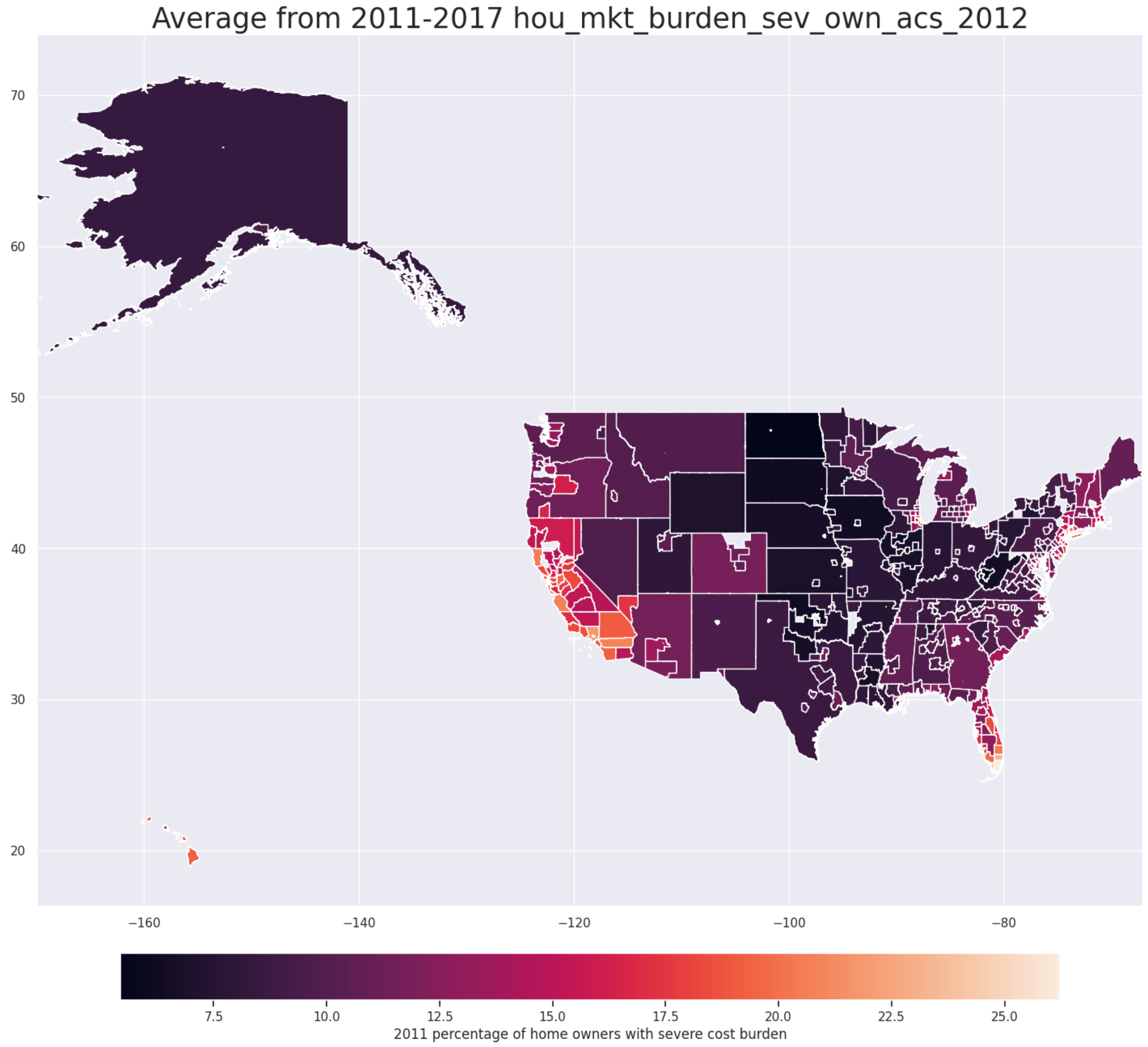
	2011)	
hou_mkt_burden_own_acs5yr_2012	2011 percentage of home owners with cost burden	Housing
hou_mkt_burden_own_acs5yr_2017	2016 percentage of home owners with cost burden	Housing
hou_mkt_burden_own_acs5yr_diff	change in hou_mkt_burden_own_acs5yr values (2016 and 2011)	Housing
hou_mkt_burden_sev_own_acs_2012	2011 percentage of home owners with severe cost burden	Housing
hou_mkt_burden_sev_own_acs_2017	2016 percentage of home owners with severe cost burden	Housing
hou_mkt_burden_sev_own_acs_diff	change in hou_mkt_burden_sev_own_acs5yr values (2016 and 2011)	Housing
hou_mkt_burden_rent_acs5yr_2012	2011 percentage of renters with cost burden	Housing
hou_mkt_burden_rent_acs5yr_2017	2016 percentage of renters with cost burden	Housing
hou_mkt_burden_rent_acs5yr_diff	change in hou_mkt_burden_rent_acs5yr values (2016 and 2011)	Housing
hou_mkt_burden_sev_rent_acs_2012	2011 percentage of renters with severe cost burden	Housing
hou_mkt_burden_sev_rent_acs_2017	2016 percentage of renters with severe cost burden	Housing
hou_mkt_burden_sev_rent_acs_diff	change in hou_mkt_burden_sev_rent_acs5yr values (2016 and 2011)	Housing
hou_pol_hudunit_psh_hud_share	share of HUD-subsidized housing units to total housing units	Safety Net
hou_mkt_pmt_unit_census_share	share of new housing permits to total housing units	Housing
urban_cat	urbanicity category	Subgroup
state	numerical state ID, based on state_abr	Geography
evict_file_flag	flag for missing eviction filing rate value	Housing
evict_flag	flag for missing eviction rate value	Housing
fedfundcoc_flag	flag for missing funding value	Safety Net
hou_mkt_density_dummy	indicator for high housing density CoC (>= 75th percentile)	Housing
ln_econ_labor_medinc_census	log of median income	Economic
flag_d_hless	flag for missing total homeless share value in 2017 or 2013	Outcome
flag_xt_hless	flag for missing total homeless share value in an odd year	Outcome
flag_d_shelt	flag for missing sheltered homeless share in 2017 or 2013	Outcome
flag_xt_shelt	flag for missing sheltered homeless share value in an odd year	Outcome
flag_d_unshelt	flag for missing unsheltered homeless share in 2017 or	Outcome

	2013	
flag_xt_unshelt	flag for missing unsheltered homeless share vale in an odd year	Outcome
census_region	census region	Geography
census_division	census region	Geography
sub_west_coast	indicator for west cost CoC (CA, OR, or WA)	Subgroup
sub_low_rent_vacancy	indicator for CoCs with rental vacancy rates <= 5 percent	Subgroup
sub_high_cost_rent75	indicator for CoCs with median rents >= 75th percentile	Subgroup
sub_high_cost_homeval75	indicator for CoCs with median home value >= 75th percentile	Subgroup
sub_high_rent_share75	indicator for CoCs with share of renters >= 75th percentile	Subgroup
tight_high_cost_rental_mkt	sum of the four tight market criteria indicators	Subgroup
sub_tight_high_cost_rent	indicator for tight, high cost rental market CoCs	Subgroup
sub_west_coast_all_urb	indicator for suburban or major city/largely urban CoCs in the west region	Subgroup
sub_west_census	indicator for west region CoCs	Subgroup
major_city	indicator for major city or largely urban CoC	Subgroup
suburban	indicator for largely suburban CoC	Subgroup
rural	indicator for largely rural CoC	Subgroup

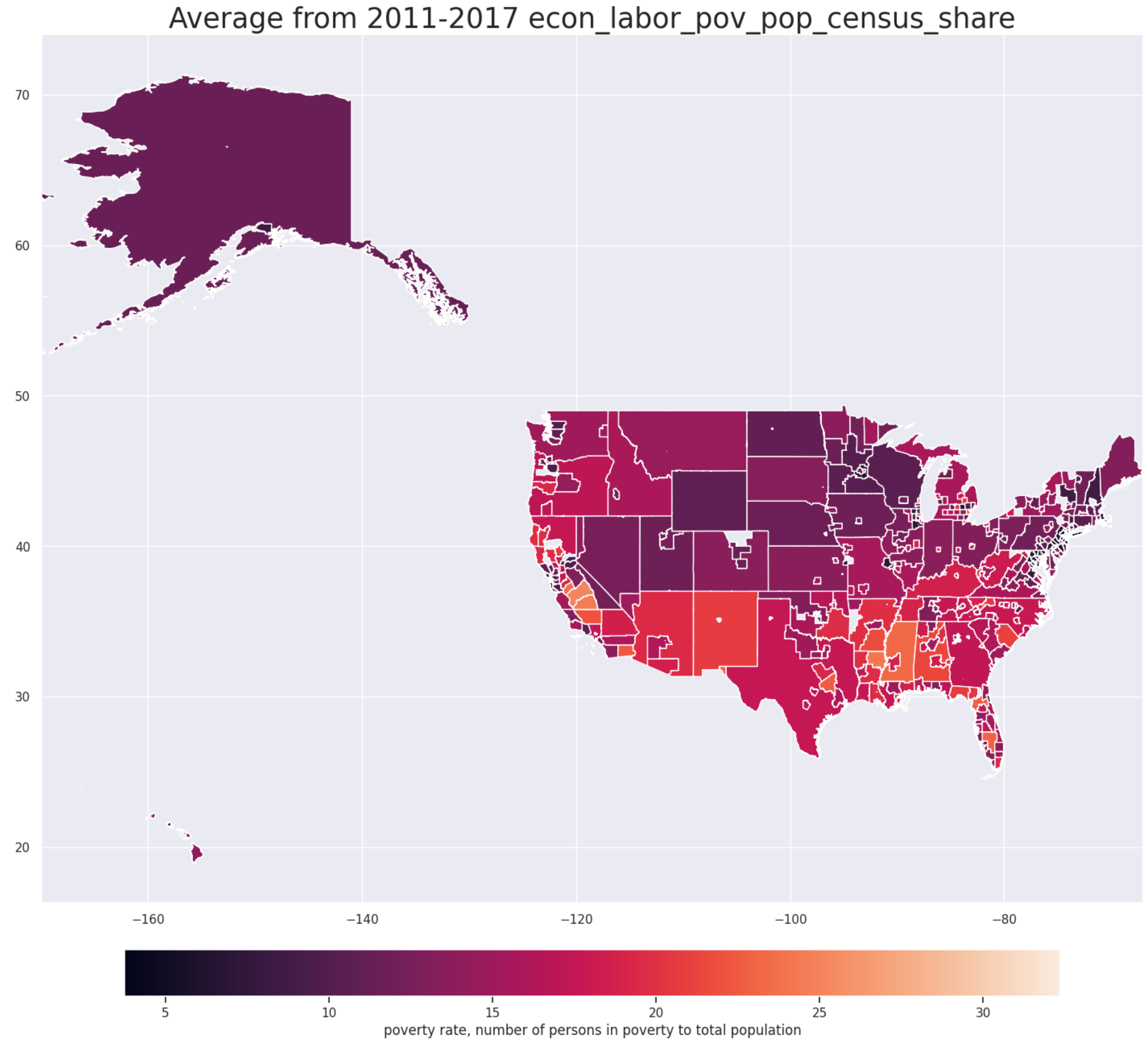
Appendix 2: Feature definitions and associated domains for all the features in the dataset after missing values were handled.



Appendix 3: A heatmap of 2011 share of renters foreach community in the U.S. The data for the color feature is averaged for each year.


Average from 2011-2017 hou_mkt_burden_sev_own_acs_2012
70
60
50
40
30
20
−160
−140
−120
−100
−80
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
2011 percentage of home owners with severe cost burden

Appendix 4: A heatmap of 2013 percentage of home ownerswith severe cost burden for each community in the U.S. The data for the color feature is averaged for each year.


Average from 2011-2017 econ_labor_pov_pop_census_share
70
60
50
40
30
20
−160
−140
−120
−100
−80
5
10
15
20
25
30
poverty rate, number of persons in poverty to total population

Appendix 5: A heatmap of the rate of poverty for eachcommunity in the U.S. The data for the color feature is averaged for each year.

Work Cited

Appelbaum, R. P., Dolny, M., Dreier, P., & Gilderbloom, J. I. (1991). Scapegoating Rent Control: Masking the Causes of Homelessness. *Journal of theAmerican Planning Association*, *57*(2), 153–164. doi:10.1080/01944369108975484

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

Burt, M. (2005, May). *Strategies for Preventing Homelessness*. Urban Institute. Retrieved December 7, 2022, from https://webarchive.urban.org/publications/1000874.html

Byrne, T., Munley, E. A., Fargo, J. D., Montgomery, A. E., & Culhane, D. P. (2013). New Perspectives on Community-Level Determinants of Homelessness. *Journal of Urban Affairs*, *35*(5), 607–625. doi:10.1111/j.1467-9906.2012.00643.x

Corinth, K., & Lucas, D. S. (2018). When warm and cold don't mix: The implications of climate for the determinants of homelessness. *Journal of HousingEconomics*, *41*, 45–56. doi:10.1016/j.jhe.2018.01.001

Culhane, D. P. (2008). The Cost of Homelessness: A Perspective from the United States. *European Journal of Homelessness,*97-114. Retrieved from https://repository.upenn.edu/spp_papers/148

Finnegan, M. (2020, March 1). *How would Democratic candidates fix the Housing and homelessness crises?* Los Angeles Times. Retrieved December 7, 2022, from https://www.latimes.com/politics/story/2020-02-18/democratic-presidential-candidates-healthcare-homelessness-policy

Grant, R., Gracy, D., Goldsmith, G., Shapiro, A., & Redlener, I. E. (2013). Twenty-Five Years of

Child and Family Homelessness: Where Are We Now? *American Journal of Public Health*, *103*(S2), e1–e10. doi:10.2105/AJPH.2013.301618

Grimes, P. W., & Chressanthis, G. A. (1997). Assessing the Effect of Rent Control on Homelessness. *Journal of Urban Economics*, *41*(1), 23–37. doi:10.1006/juec.1996.1085

Healy, J. (2013, January 12). *In Wyoming, Many Jobs but No Place to Call Home*. Retrieved from https://www.nytimes.com/2013/01/13/us/homelessness-increases-in-wyoming-product-of-economic-boom.html

Nasar, H., Vachon, M., Horseman, C., & Murdoch, J. (2019). Market Predictors of Homelessness: How Housing and Community Factors Shape Homelessness Rates Within Continuums of Care | HUD USER. Retrieved December 7, 2022, from https://www.huduser.gov/portal/publications/Market-Predictors-of-Homelessness.html

Number of renter occupied homes in the U.S. 2021. (n.d.). Retrieved 7 December 2022, from Statista website: https://www.statista.com/statistics/187577/housing-units-occupied-by-renter-in-the-us-since-1975/.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Toro, P. A., Trickett, E. J., Wall, D. D., & Salem, D. A. (1991). Homelessness in the United States: An ecological perspective. *American Psychologist, 46*(11), 1208–1218. https://doi.org/10.1037/0003-066X.46.11.1208

Vorotyntsev, D. (2020, September 12). *Stop permuting features*. Medium. Retrieved December 8, 2022, from https://towardsdatascience.com/stop-permuting-features-c1412e31b63f

Williams, D. (2021, January 8). *U.S. homes for sale are getting more expensive despite the coronavirus*. Forbes. Retrieved December 7, 2022, from https://www.forbes.com/sites/dimawilliams/2020/05/21/us-homes-for-sale-are-getting-more-expensive-despite-the-coronavirus/