Understanding Homelessness: A Comprehensive Analysis of Economic and Housing Factors in Urban America

Jasmine Jia

HOMELESSNESS CRISIS IN THE U.S.

- As of 2023, there are an estimated **580,000 homeless individuals** in the U.S., with over 60% residing in shelters or temporary housing and approximately 40% unsheltered. Rising housing costs, income inequality, and the COVID-19 pandemic have intensified the crisis, highlighting the need for policy interventions.

OBJECTIVE

This project employs machine learning to **predict homelessness rates across the U.S.**, aiming to provide data analytics insights that help policymakers craft targeted, sustainable interventions.

Through this approach, we seek to enhance the understanding of how **different societal factors influence homelessness and to support the development of effective strategies to combat this pressing social issue.**

DATA & METHODOLOGY

Data Source

The dataset involves an annual Point-in-Time (PIT) Count to measure homelessness **on a specific night,** encompassing **sheltered** and **unsheltered** populations, while the Housing Inventory Count (HIC) catalogs available beds and units dedicated to those experiencing homelessness. Both datasets are submitted to HUD via the Homelessness Data Exchange (HDX), enabling comprehensive analysis and informed policy-making.

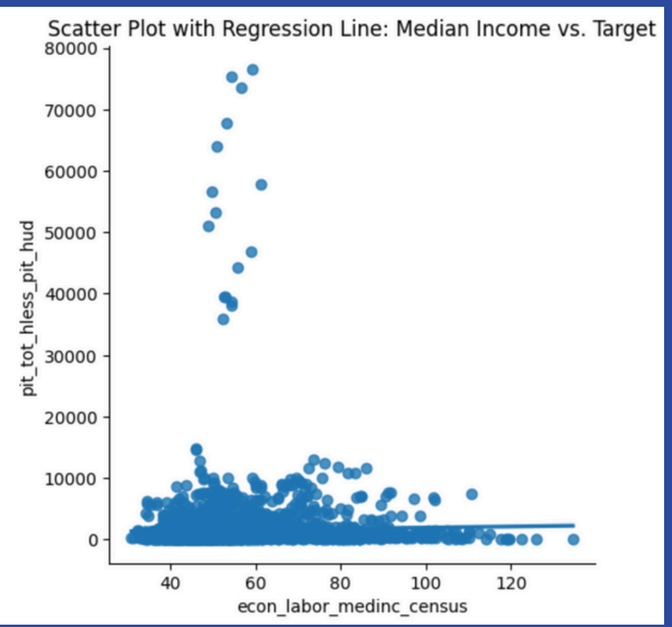
Model Selection

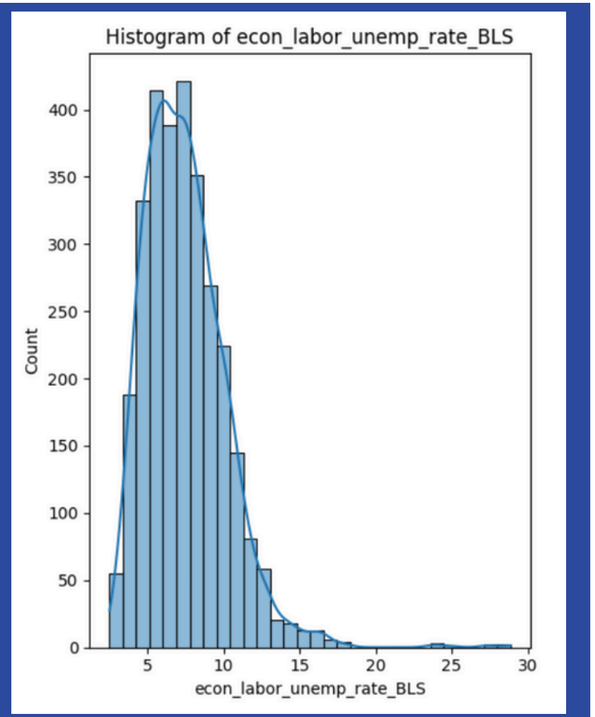
The model aims at understanding homelessness in the U.S., we selected a wide range of **socio-economic, housing, and demographic variables** that significantly impact **homelessness population**.

	coef	std err	t	P>|t|	[0.025	0.975]
const	−418.8149	452.008	−0.927	0.354	−1305.182	467.552
econ_labor_medinc_census	−31.0430	6.742	−4.605	0.000	−44.263	−17.823
econ_sn_eitc_irs_share	11.6266	11.811	0.984	0.325	−11.535	34.788
econ_labor_unemp_rate_BLS	67.1197	17.159	3.912	0.000	33.472	100.767
hou_mkt_medrent_acs5yr_2017	41.9113	48.433	0.865	0.387	−53.063	136.885
hou_mkt_homeval_acs5yr_2017	3.8343	0.887	4.323	0.000	2.095	5.574
hou_pol_bed_psh_hic_hud	1.9106	0.034	56.584	0.000	1.844	1.977
dem_pop_density_census	3.3718	2.560	1.317	0.188	−1.649	8.393
dem_soc_white_census	0.0005	7.66e−05	6.403	0.000	0.000	0.001
dem_soc_black_census	0.0009	0.000	2.939	0.003	0.000	0.002

Among all the independent variables, **income level, unemployment rate, house values, numbers of pernament supportive beds, and racial factors** (White or Black) are statitically significant.

MODEL EVALUATION

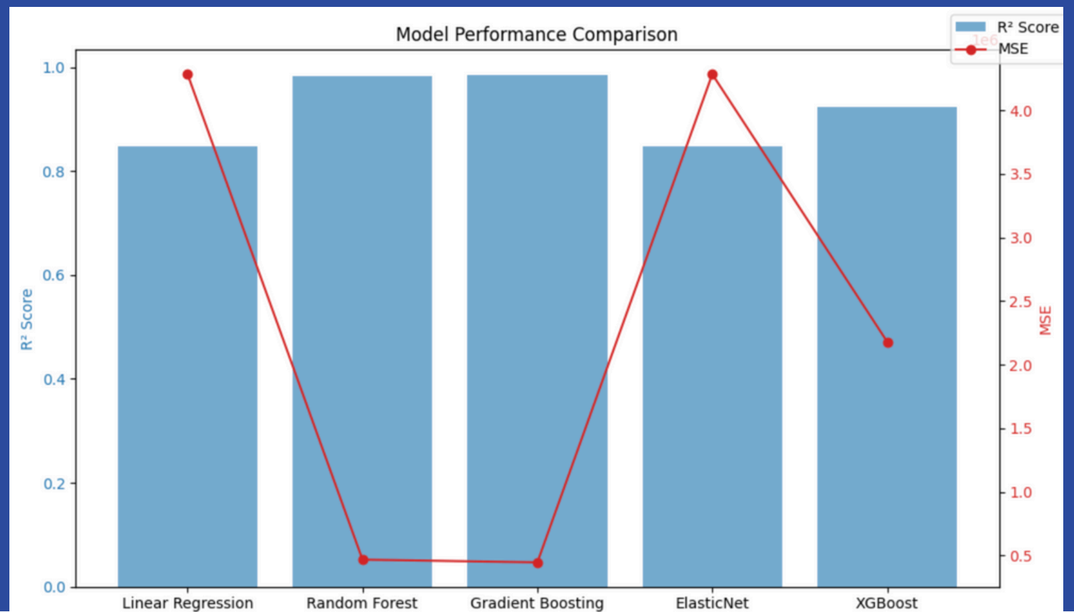




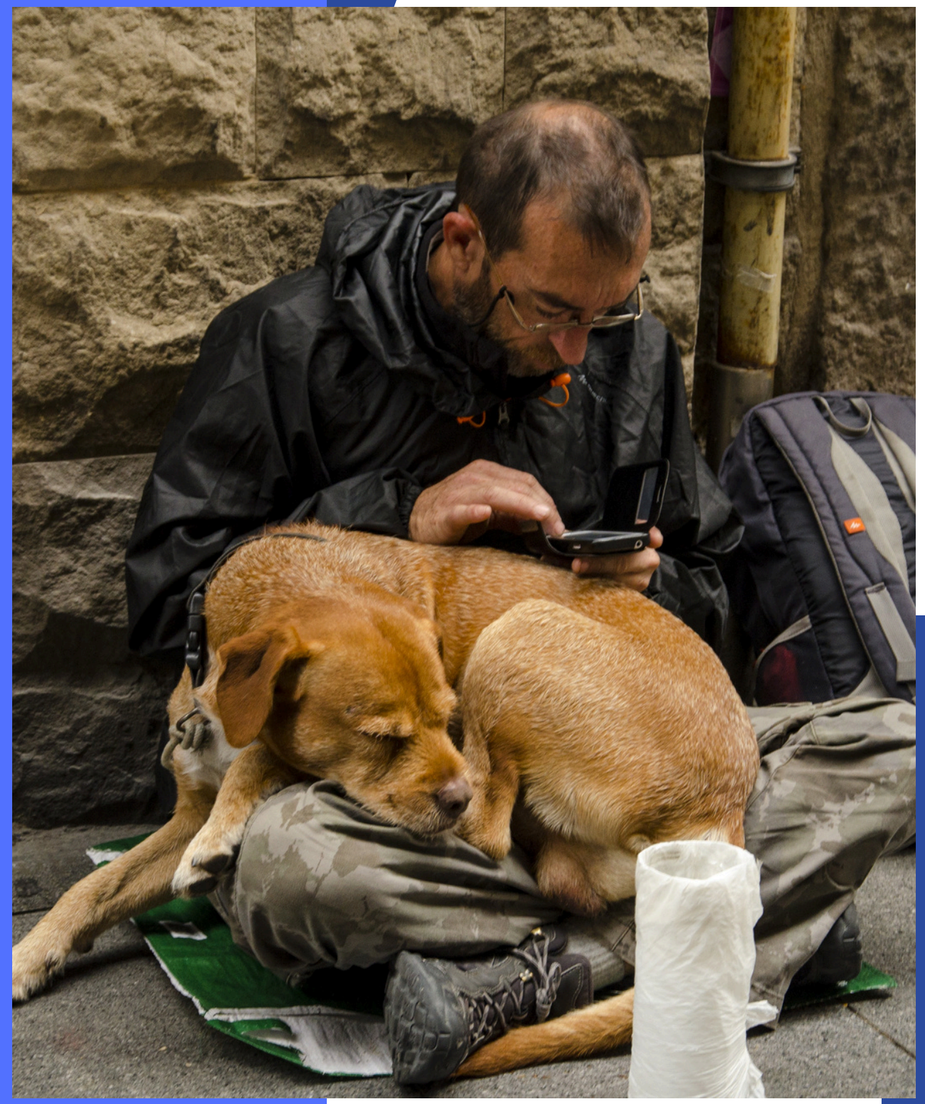
The **scatter plot** with a regression line reveals a **non-linear** relationship between median income and total homelessness population, suggesting that linear regression may not accurately capture this relationship. Therefore, tree-based models, which work well in handling non-linear patterns, are a more suitable choice. The **histogram** demonstrates that areas with higher unemployment rates tend to have a higher count of individuals experiencing homelessness, suggesting a positive correlation between these variables. The skewed distribution emphasizes how economic downturns impact vulnerable populations the most, leading to increased homelessness in affected regions.

```
Linear Regression: R² = 0.8477, MSE = 4282419.4383
Random Forest: R² = 0.9834, MSE = 466823.5745
Gradient Boosting: R² = 0.9842, MSE = 445394.8065
ElasticNet: R² = 0.8477, MSE = 4283359.0089
XGBoost: R² = 0.9225, MSE = 2180417.7801
```

Random Forest and **Gradient Boosting** provided the best performance with R^2 scores above 0.98 and significantly lower MSEs, underscoring their ability to handle non-linear relationships effectively and to capture intricate patterns in the data without overfitting. The GBM model, provids competitive results, offering insights into the stability and reliability of using boosted trees for this type of data.



LIMITATION



1. **Limited Racial Representation:** The focus on Black and White populations excludes other racial and ethnic groups, limiting insights into the broader homelessness issue.
2. **Point-in-Time Data Not Accurate**: PIT counts, offering a one-night snapshot, can underrepresent the homeless population due to undercounting and the dynamic nature of homelessness.
3. **Model Complexity and Overfitting**: Complex models may overfit to training data, reducing their ability to generalize and accurately predict across various datasets.
4. **Limited Feature Diversity:** The dataset lacks features covering all dimensions of homelessness, restricting the model's predictive capability.

NEXT STEPS

1. **Expand Racial Representation**: Collect more data on different racial and ethnic groups to gain comprehensive insights into homelessness.
2. **Improve Data Collection Accuracy**: Implement continuous data collection methods to supplement PIT counts, achieving more accurate data.
3. **Simplify Model Complexity:** Use simpler models or regularization to prevent overfitting and improve generalization.
4. **Enhance Feature Diversity:** Add features reflecting various factors influencing homelessness for improved model predictions.