Summer
2025

White Paper: Regional Childcare Affordability Indexing & Prediction

DSC680-T301

LUKE SYVERSON

Topic

The “Regional Childcare Affordability Indexing & Prediction” project will seek to measure and predict the relative affordability of childcare across many regions of the United States. The ambition is to enrich the analysis surrounding childcare affordability and to provide policy makers insights on contributing factors to rising childcare costs in their respective regions in the US.

Business Problem

Childcare costs are a complex problem for many families in the US. Childcare deemed “unaffordable” by a lower-income segment of the population means that a larger portion of the potential workforce would either choose at-home childcare instead of a full-time job, possibly reducing gross workforce productivity in the short term, or suffer a smaller discretionary income despite pay from a full-time job, especially in the case of single parents. If policy makers can alleviate the financially based difficulties from unaffordable childcare, more people with children in the lower-income brackets will find more time to pursue full-time employment or retain more of their income to spend on other essential costs. The premise, therefore, is that one of the largest burdens on the relatively impoverished population in the US are parents in lower-income brackets with little-to-no means to afford childcare, which impacts regional and national productivity, but may be alleviated by means of fiscal policy enacted to minimize the impact.

Background

Childcare costs have continued to climb over the past decade, placing growing pressure on families—particularly those in low- to moderate-income brackets. As wages stagnate in many sectors and the cost of living rises, the burden of securing affordable childcare has become a critical concern for working parents. While national trends provide a broad view of the issue, the affordability of childcare varies significantly from one region to another, shaped by differences in local economies, labor markets, and policy environments. This project recognizes that understanding regional nuances is essential for crafting effective, data-driven solutions.

Preliminary Analysis

Judging by the feature labels and the technical guide, most of the features are aggregates or proportions relative to meaningful subsets of the population. Percentiles attempt to simulate differences between socioeconomic circumstances.

The data is continuous from 2008 to 2022, meaning 14 years of data from most US counties were collected; however, analysis has yet to determine which states and counties are underrepresented in the dataset. Missing value analysis should be conducted on the features to determine the relative coverage of counties within states, and where the predictive power may skew towards.

As the grain is at the county-year level, counties should be represented equally in the model, but data collection practices across counties are expected to vary in quality due to reporting limitations and possible incomplete yearly coverage of prices.

Methods

The analysis methods include an automated feature selection to trim the feature set to the most powerful predictors available, to then be utilized in a gradient-boosted model to predict an engineered "affordability index" that seeks to compartmentalize the factors of "affordable childcare" into an adjustable target field for the model.

Analysis

The feature selection and transformation pipelines behind the predictive model are tiresome, but parameterized and functioning. Analysis leveraging these pipelines creates a slim model with overfitting tendencies and possible multicollinearity, but the predictions indicate potential to maximize the information gained from noise surrounding the affordability construction.

Regarding the data model, visualizations in Power BI reveal variance across counties and states that warrants future investigation, possibly by a less-generalized model.

Figure 1 below displays a sample disparity between states where affordability vs. median income is charted, but population isn't controlled for.

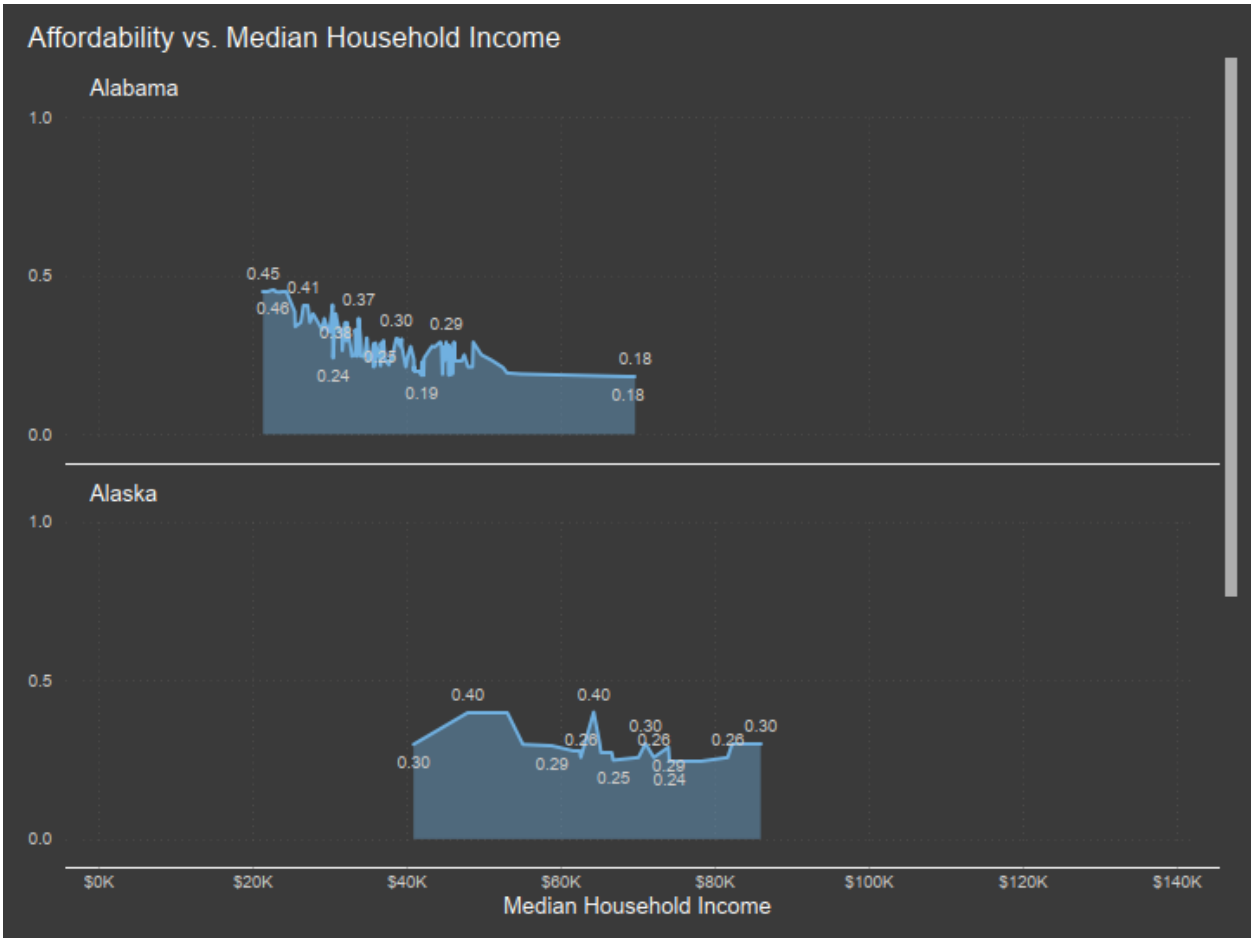

Figure 1

Conclusion

The sensitivity and complexity of childcare costs warrant a thoroughly vetted model that hedges against bias and overfitting to provide conservative claims for policy analysis, as superfluous claims or bias would negatively influence policy decisions and potentially induce the undesired effect of increasing childcare costs, which further burdens economies and struggling individuals. The current model requires additional tuning to minimize overfitting and provide more helpful predictions for a policy-making foundation.

Assumptions & Limitations

Much of this data is numeric, presumably comprised of percentiles, ratios, flags, and sums. Filtering predictors and encoding flags for analysis will be crucial to make sense of this in future analysis.

The technical guide implicates different collection policies from different states over the lengthy duration of data collection, meaning noise may be present when comparing different states and subsequent collection practices.

Figure two displays a percentile-based approach to childcare costs over a 10-year range. Notably, the difference between 75th percentile and the median are minimally informative for all states and begs the question of its inclusion in the data model for feature selection.

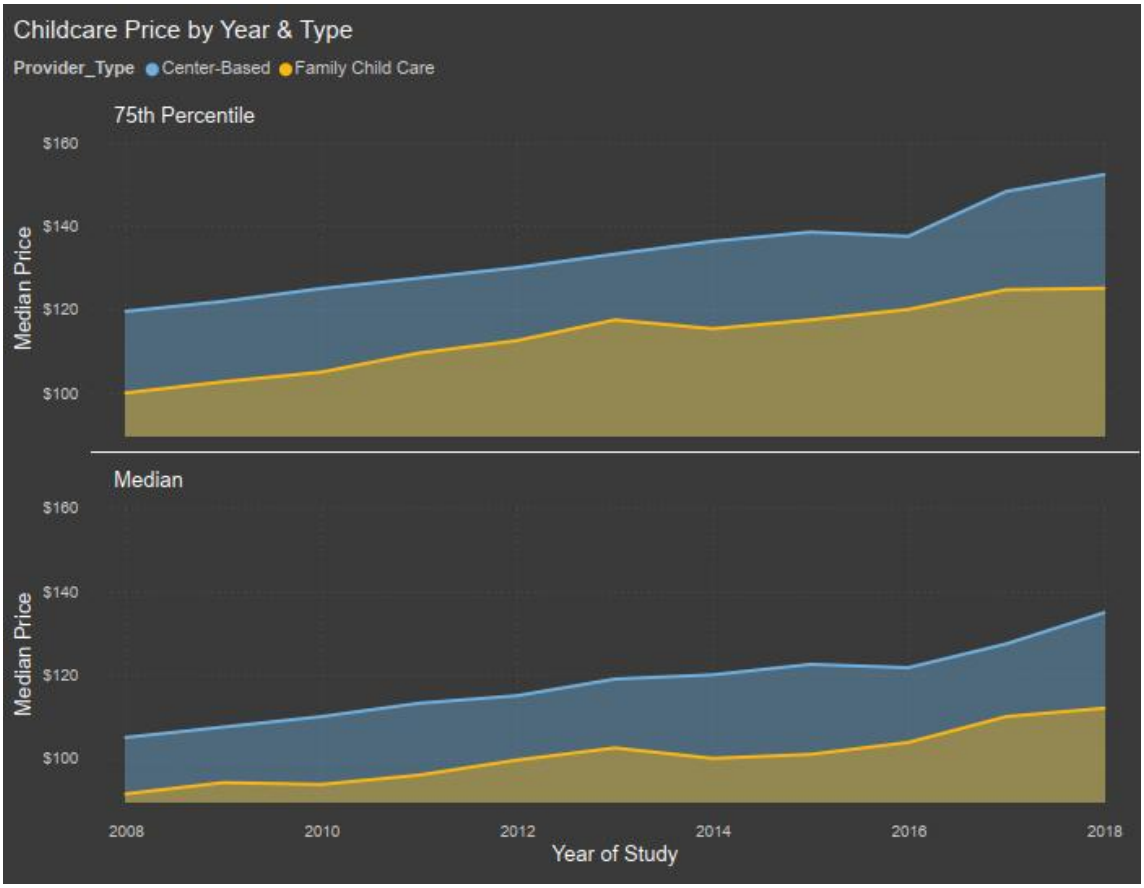

Figure 2

Challenges

The sheer number of features is daunting and requires an attuned analysis to successfully reduce the data to a predictable state. Given the breadth of time the data are collected over, the model should be resistant to missing values (else excluded) and able to predict in a generalizable fashion the trends impacting various regions. It's possible the effects of predictable childcare vary across regions, in which case, the regional data should be engineered to indicate broader regions of the US instead of State and County-level.

Figure three indicates the density of the model by displaying the clusters of county names, grouped by state, alongside their average affordability index. Aggregation methodologies and possibly feature enrichment (via normalized data modeling for geographic data) could simplify the difficulty of parsing so many observations.

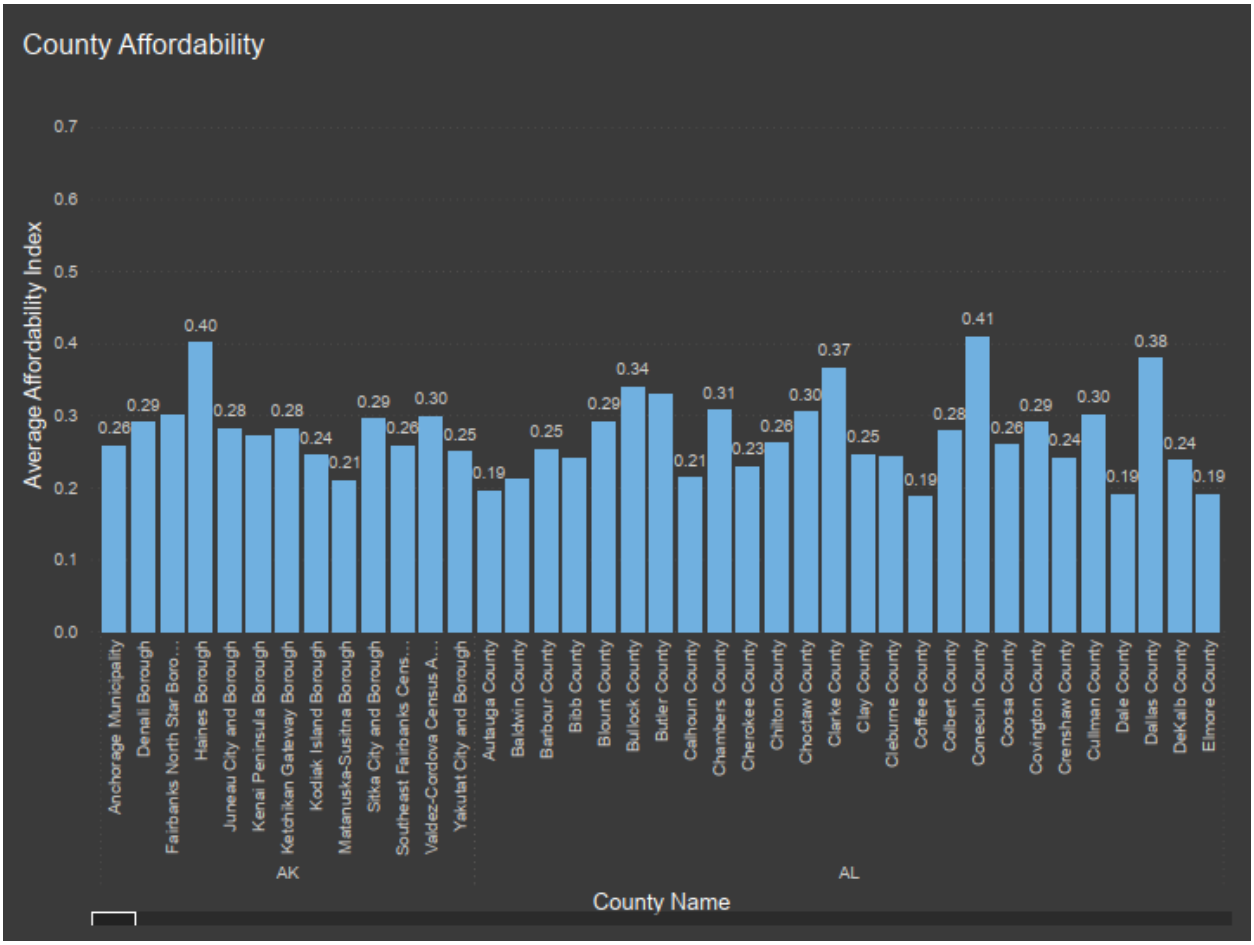

Figure 3

Future Usage & Implementation

The predictive model may be used by policy making aids to advise policy groups of strategies to inspire changes in regional childcare affordability. The layered approach of the analysis that provides an interactive component via Power BI and a more traditional notebook for data pipelines for engineering and modeling will service the different technical competency levels of interested parties.

Sample visualizations against a template, star schema model will be provided in Power BI to inspire future analysis. The PBI model may not be suitable for modeling predictions, depending on the nature of the chosen model and the relative interpretability of the results.

Recommendations

A simplified dataset should be provided that engineers interaction and aggregation features to consolidate the massive data model. Generally feature scarcity is an issue, but navigating dozens of permutations of the same feature doesn't capture the relationships between features and only serves to obfuscate feature selection methods.

Models should undergo more thorough cross validation and predict more aspects of the data model than just affordability parameters. The potential to unearth demographic trends adjacent to childcare cost influencers may also help policy makers exercise targeted initiatives to alleviate specific regional burdens for overrepresented childcare burdens within demographics.

Ethical Assessment

Relating numerical predictions to the difficulties of affordable childcare is an arduous task that requires creativity in presentation and the utmost integrity in statistical representation. A skewed analysis is ethically unacceptable, as conclusions drawn from inaccurate presentations risk the increase of financial burden upon struggling families by policy application that favors groups with negligible difficulties in procuring childcare. Accurately forecasting prices allows efficient budgetary allocation to subsidies and other economic stimuli that effectively alleviate financial burden upon low-income groups.

Sample Questions

1. How could a predictive model help those burdened with unaffordable childcare?
2. How could time spent working with data be more impactful than volunteering?
3. How can policy aides use an interactive report to better understand their regions?
4. What information can be retrieved from a Power BI report to help inform policy making?
5. What makes an affordability index helpful?
6. Why not predict other aspects about childcare than an abstract affordability measure?
7. How is inflation controlled for in the affordability index?
8. How are different costs of living accounted for in the affordability index?
9. Where can an affordability index be inaccurate?
10. Why would an affordability index be inaccurate?

Appendix & References

ICF. (2020). National Database of Childcare Prices: Final Report. In *U.S. Department of Labor* (pp. 1–94). U.S. Department of Labor, Women's Bureau. www.dol.gov/agencies/wb/topics/featured-childcare

RegionTrack, Inc. (30 C.E.). Child Care in State Economies [Review of *Child Care in State Economies*]. In *The Conference Board* (pp. 1–65). Committee for Economic Development. https://www.ced.org/assets/reports/childcareimpact/181104%20CCSE%20Report%20Jan30.pdf