How Does Metropolitan Sustainability Efforts Influence Housing Prices in the Greater Washington Area?

Josiah Valdez & Shiyu Xu

July 2, 2024

Abstract

This project examined the relationships between property values, walkability, electric vehicle (EV) infrastructure, and the sustainability profile of residential areas in Washington state. Utilizing a comprehensive methodology that includes regression analysis, classification models, and machine learning, we intend to uncover how sustainability factors influence real estate values. By integrating datasets from Zillow, the US EPA National Walkability Index, and EV registrations in the state of Washington, we were able to gain a nuanced understanding of the economic impacts of sustainable living on the real estate market.

1 Introduction

As cities evolve, the importance of sustainable urban development becomes increasingly evident. Factors such as walkability and access to electrical vehicle infrastructure are not only vital for environmental sustainability but may also have a significant impact on property values. This project explores these relationships within the state of Washington, offering insights into how sustainability factors correlate with and potentially influence real estate values. Through a multi-methodological approach, By using various Machine Learning methods, we categorized each zip code in Washington State into assigned sustainability labels of High Sustainability & High Growth, Low Sustainability & Declining Markets, and Emerging Sustainability Markets. Ultimately, this research aims to understand and contribute to the discourse on urban planning and real estate development.

2 Methods

2.1 Data Engineering:

The data used in this project is sourced from publicly available government and company databases. The Zillow Research page provides public data on home value, forecast, rental, and property sales across different regions within the U.S., and we utilized the Home Values and Home Values Forecasts datasets from Zillow as they were the best indicators of current housing values in the market. The Walkability Index dataset documents relative walkability for every 2019 Census block group and provides a national

Walkability Index score, which is be used to explore connections to the property values.

We aligned the Zillow Home Value Index, the Zillow Home Value Forecast, the EPA National Walkability Index, and the Electric Vehicle Population datasets based on geographical and temporal parameters to create a unified database for analysis. We used US zip codes as the ID to which all datasets are mapped to. To convert features taken from the National Walkability Index dataset to be zipcode-based, we incorporated a table from the USPS that details which US census tract groups belong to which zip codes. The EPA database contained some census tract numbers that were not found in the USPS database because of their block numbers, so we worked under the assumption of categorizing them with the greater tract number (BLKGRPCE = 00) in terms of the relevant features.

Note about the US Census tract number: a census tract can be in more than one zip code and zip codes can contain several census tracts. When we map the tract values to the zip code dataset, a census tract will get mapped to only one zip code even if there is a many-to-many relationship. We think this is fine for our intents and purposes as we can't ascertain the certain percentage of the census tract population that resides in one part of the zip code or the other. This also prevents double-dipping when building the relationship to the zip codes.

2.2 Regression Analysis:

To assess the correlation of the walkability scores, EV population, and other features with property value, we created a correlation matrix and graphical visualizations to get a better sense of any feature correlation that may exist (Figure 1).

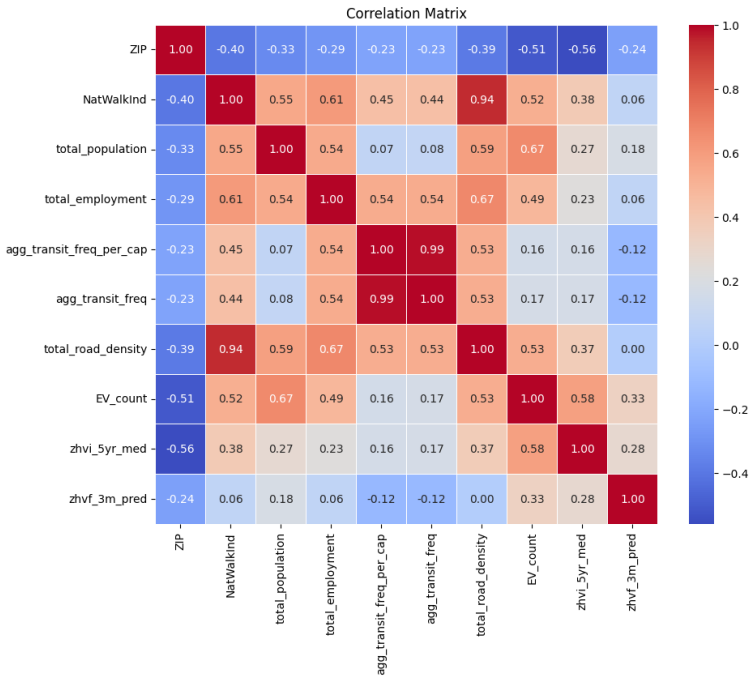

Figure 1: Correlation Matrix of all features

The higher correlation observed between some features leads us to the decision of

prediction with random forests, due to the algorithm's random features subsets and bootstrapping design that ideally reduces the model over-fitting to our data.

2.3 Classification:

Other than building a regression model, we created the classes "High Sustainability & High Growth Market", "Low Sustainability & Declining Market", and "Emerging Sustainability Market" to categorize the zip codes in Washington. To create the classes, we first normalized the data and then created a weighted sustainability score, S, in which the dataset is re-evaluated.

$$S = \sum_{i=1}^{N} w_i \cdot f_i$$

Where f_i represents the value of the i-th feature, w_i is the weight assigned to the i-th feature, and N is the total number of features. For our purposes, the weights for each feature (f_i) are set equally, but further research can be done to calculate the exact weight each feature should have on the scoring process.

The classes are determined by binning the zip codes into high, medium, and low growth categories based on the top, middle, and bottom tertiles of the Zillow Home Value Forecast (ZHVF) values; as well as binning by each tertile for their sustainability scoring S. From here, we label the highest third in both ZHVF and S score as High Sustainability & High Growth Market, the lowest third in both fields as Low Sustainability and Declining Market, and the rest as Emerging Market. (Figure 2)

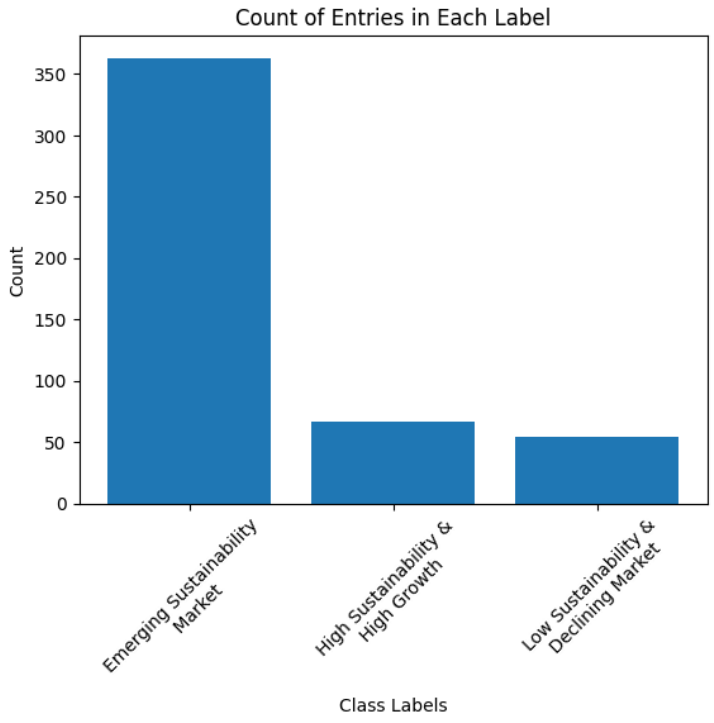

Figure 2: Distribution of Market Class Labels

To implement the classification model, we chose to use the XGBoost Classifier, as the boosting algorithm would be able to capture more nonlinear relationships. To improve the accuracy of the model, we also experimented with various depths and cross-validation.

3 Experiments and Results

3.1 Data Exploration

We employed linear regression to assess the direct correlation of the Walkability scores, EV population, and other sustainability metrics such as distance to availability on property values.

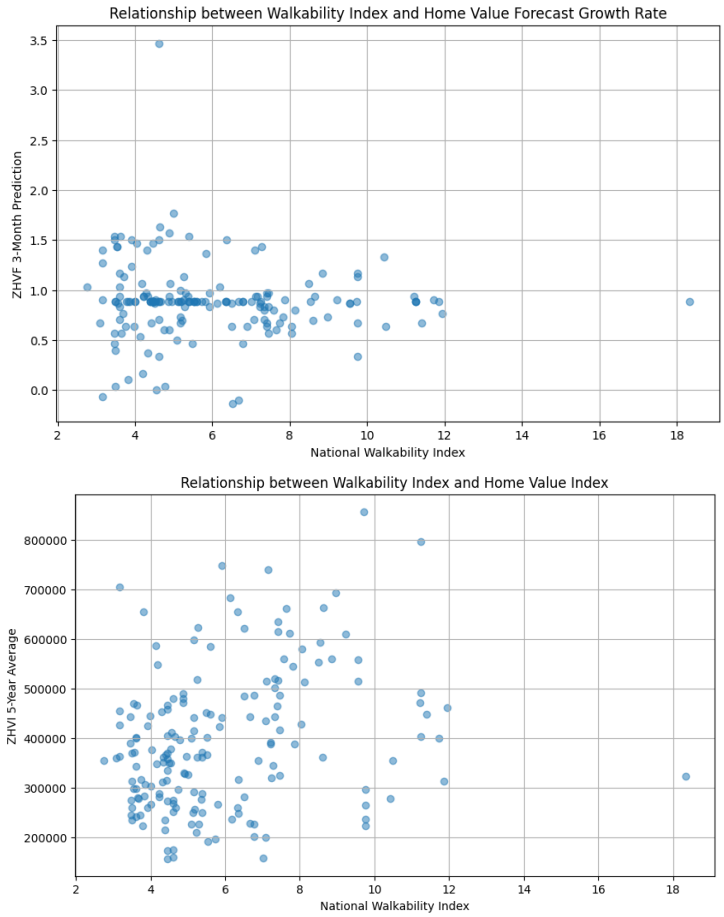

Figure 3: Relationship of property values to Walkability Index

Before this investigation, we had naively hypothesized that the Walkability index would be the best predictor for housing prices. From these plots (Figures 3, 4), we can observe that the relationship is not as direct as we predicted, in comparison to the relationship of property values to some of the other features in this dataset:

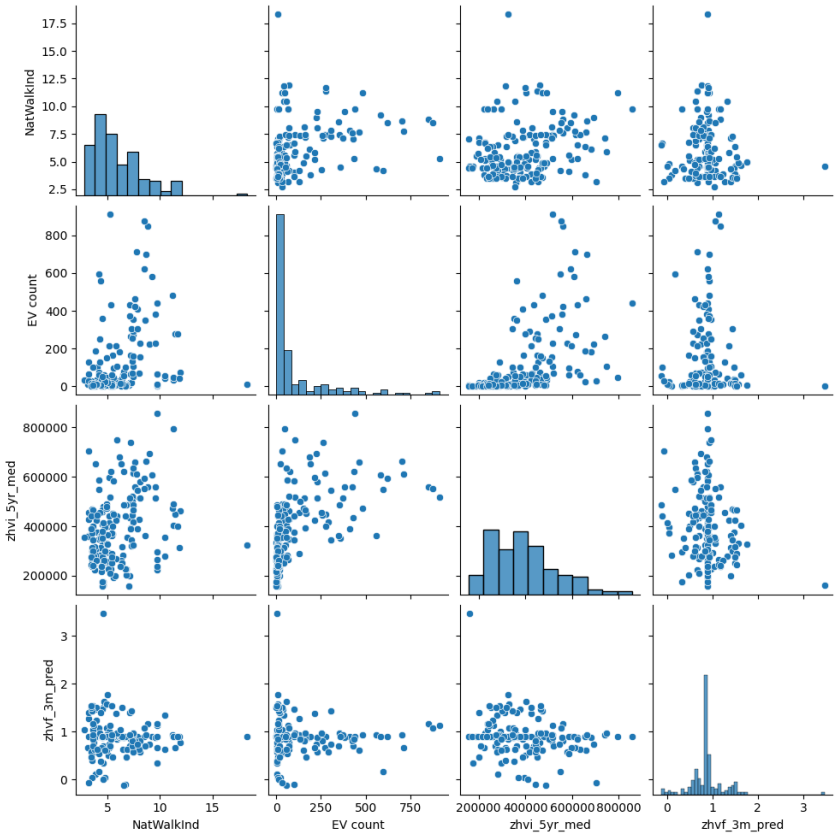

Figure 4: Relationship Plots for Walkability Index, EV Count, ZHVI, ZHVF

3.2 Regression

We began by building a regression model from our data with multiple linear regression, predicting the median Home Value Index (zhvi_5yr_med). To evaluate this model, we calculated its various error rates, as well as plotted the true versus predicted values for our test set (Figure 5).

The regression model had a mean squared error (MSE) $\approx$ 27850000000 due to the size of the data point values, a root mean squared error (RMSE) $\approx$ 166900, a mean absolute error (MAE) $\approx$ 122600 and a normalized MSE $\approx$ 0.01711. The r^2 score of this model was $\approx$ 0.5812.

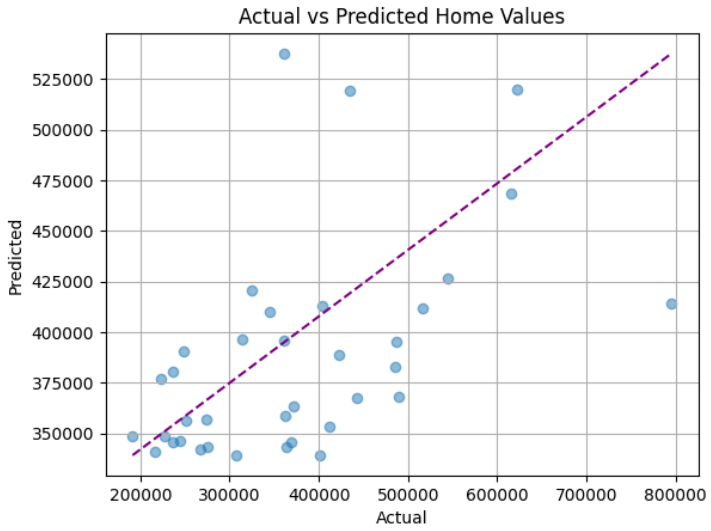

Figure 5: Actual vs. Predicted Home Values (ZHVI) for Linear Regression Model

We had anticipated the low r^2 scoring as a result of the dataset's complex relationships, so we built a random forest model to better capture relationships during the prediction process (Figure 5).

The random forest regression (Figure 6) showed better results, with MSE $\approx$ 16480000000, MAE $\approx$ 90680, RMSE $\approx$ 128400, and normalized MSE $\approx$ 0.01013. The r^2 score of the actual versus predicted values was $\approx$ 0.7521. The improvements in prediction error rates may be due to random forests being less prone to overfitting, which in turn gives better predictions on the unseen test sets.

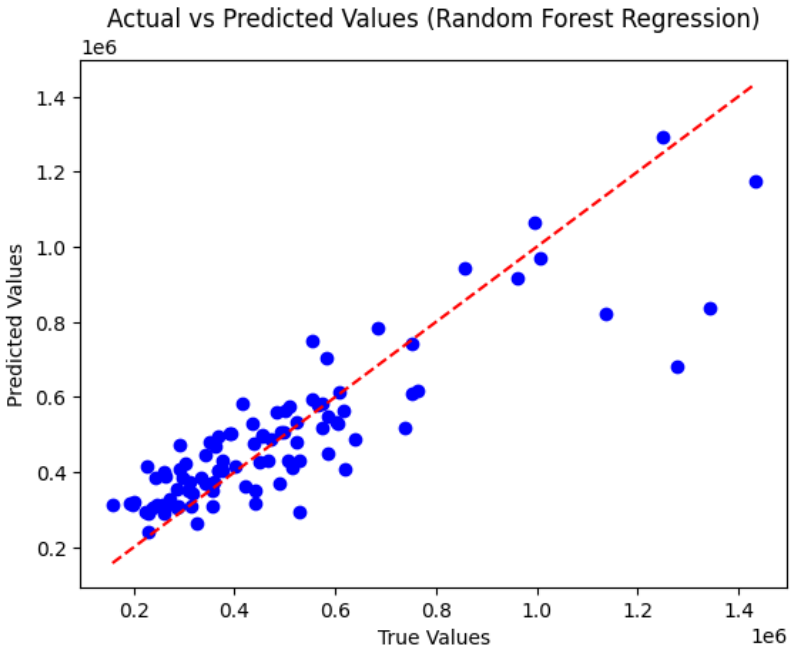

Figure 6: Actual vs. Predicted Home Values (ZHVI) for Random Forest Regression

The random forest regressor was also able to give insight into what feature(s) was most significant in the regression process:

1. EV_count
2. agg_transit_freq
3. zhvf_3m_pred
4. total_employment
5. total_population
6. agg_transit_freq_per_cap
7. total_road_density
8. NatWalkInd

This contradicted our initial hypothesis that the walkability index (NatWalkInd) would have the most feature importance. While EV count as the most significant feature is logical given the financial requirements of owning an electric vehicle, it was surprising that total employment in an area was also relatively important as a feature to split on for predicting housing value, with a feature importance of ≈ 0.1229 compared to the walkability index's ≈ 0.02890.

3.3 Sustainability Scoring and Zip Code Classification

Using XGBoost's classifier, we were able to train the classification model and predict with relatively high accuracy on unseen test data. After passing in the dataset classified on sustainability scoring and potential growth (ZHVI), the classifier with native max depth of 6 was able to predict using the entire dataset with an overall accuracy ≈ 0.9072, and ≈ 0.9588 when predicting from just the sustainability scoring and the ZHVF (Figure 7).

```
Predicting from:  all features
Current max depth:  6
              precision    recall  f1-score   support

           0       0.97      0.91      0.94        79
           1       0.64      1.00      0.78         9
           2       0.78      0.78      0.78         9

    accuracy                           0.91        97
   macro avg       0.80      0.90      0.83        97
weighted avg       0.92      0.91      0.91        97

Accuracy: 0.9072164948453608
Predicting from:  ['sustain_score', 'zhvf_3m_pred']
Current max depth:  6
              precision    recall  f1-score   support

           0       1.00      0.95      0.97        79
           1       0.82      1.00      0.90         9
           2       0.82      1.00      0.90         9

    accuracy                           0.96        97
   macro avg       0.88      0.98      0.92        97
weighted avg       0.97      0.96      0.96        97

Accuracy: 0.9587628865979382
```

Figure 7: Classification Reports for XGBoost Classifier

We further experimented with different depths from three to eleven for the boosting model and found that the classifier with the maximum depth of five exhibited the best overall results when learning from the entire dataset (Figure 8). The accuracy did not change with maximum depth when predicting using just the sustainability score and the ZHVF value, remaining at ≈ 0.9588 due to the sparseness of the trees. Since the sustainability score is already calculated from the most significant features, the higher accuracy can be attributed to de-noising the model by pruning less relevant features.

```
Current max depth:  5
              precision    recall  f1-score   support

           0       0.99      0.92      0.95        79
           1       0.69      1.00      0.82         9
           2       0.80      0.89      0.84         9

    accuracy                           0.93        97
   macro avg       0.83      0.94      0.87        97
weighted avg       0.94      0.93      0.93        97

Accuracy: 0.9278350515463918
```

Figure 8: Classification Reports for XGBoost Classifier with $max\_depth = 5$

As a further experiment, we performed cross-validation on the classifier with the entire dataset and received an overall accuracy score of ≈ 0.8925 for five-fold cross-validation. In general, the XGBoost algorithm was able to classify the target categories with high accuracy at the maximum depth of 5 for the decision stumps.

4 Summary

This project comprehensively examined how sustainability factors such as walkability and electric vehicle infrastructure, as well as additional factors of population, employment, public transit accessibility, and total road density, have a nuanced impact on property values in Washington state. The regression analysis gives insight into how different environmental factors affect real estate value: from our study, we can synthesize that the EV population, transit frequency, and demographics statistics are the most influential, while the walkability index (NatWalkInd) was not as significant in the prediction process for housing prices. This could suggest that other factors associated with walkability, such as neighborhood development and amenities play intermediary roles that were not directly captured in the walkability index. So while there is a positive correlation between these factors and property values, the relationship is not straightforward or uniformly strong across all metrics. The significant impact of EV counts on property value suggests that areas with higher adoption rates of electric vehicles may be seen as more desirable, due to a higher standard of living or a greater commitment to sustainability.

For real-world applications, the labeling and classification algorithms would be useful given the necessary data to evaluate future real estate potential. These findings offer valuable insight for urban planners and policymakers focusing on sustainable development, as understanding the variables that influence property values, more targeted strategies can be developed to enhance urban living standards while boosting real estate markets.

Investors and developers can use these insights to identify potential high-growth areas, focusing on sustainability factors that are likely to attract buyers and renters. The weights for scoring sustainability can be adjusted based on need or preference for any features. As machine learning becomes increasingly applicable to every factor of life, this project serves as an example of how ML applies to the urban sciences and real estate markets.

5 References

Datasets and sources used:

- Department of Licensing. (2023). *Walkability Index* [Data set]. `https://catalog.data.gov/dataset/walkability-index1`
- US EPA. (2023). *Smart Location Database Technical Documentation and User Guide*. `https://www.epa.gov/sites/default/files/2021-06/documents/epa_sld_3.0_technicaldocumentationuserguide_may2021.pdf`
- Department of Licensing. (2024). *Electric Vehicle Population Data* [Data set]. `https://catalog.data.gov/dataset/electric-vehicle-population-data`
- Zillow Group Inc. (2024). *Housing Data* [Data set]. `https://www.zillow.com/research/data/`
- Environmental Protection Agency. (n.d.). Smart Location Mapping. EPA. `https://www.epa.gov/smartgrowth/smart-location-mapping`
- Zillow Launches New Neural Zestimate, Yielding Major Accuracy Gains. Zillow MediaRoom. (n.d.). `https://zillow.mediaroom.com`
- Office of Policy Development and Research. (n.d.). HUD USPS Crosswalk Files. `https://www.huduser.gov/apps/public/uspscrosswalk/home`