|  |
| --- |
| Haystacks.ai |
| The Convenience Factor |
| Residential Valuation Using Machine Learning |
|  |
| **Daniel Immediato** |
|  |

|  |
| --- |
| ***Abstract****: This research paper investigates the impact of Points of Interest (POIs) on residential property valuation using machine learning techniques. Specifically, it examines whether metrics such as proximity and density to external POI significantly influence property values. Grocery stores, a type of shopping center typically dependent on location, were chosen as the focal POI. Focusing on the Atlanta, Georgia Metropolitan Region, this study finds that distance and density metrics are not only significant but also substantive in determining property values. Additionally, this paper explores the types of grocery stores and identifies which factors most significantly impact valuations. The advantages and limitations of various machine learning methods are discussed, providing insights into effective feature engineering strategies for real estate valuation models.* |

**Introduction**

**Background: Property Valuation Impacted by External Factors**

What determines a property’s value? Conventional wisdom often attributes it to market dynamics—ever-changing conditions influenced by supply and demand, interest rates, and economic stability. However, property valuation, or the assessment of a property’s value, is far more nuanced, involving a combination of factors, each with varying degrees of impact. Internal factors, such as square footage, age, condition, and features like the number of bedrooms and bathrooms, play a significant role in determining a property’s price. For instance, square footage is considered a baseline feature for property value. It can generally be observed that as square footage increases, so does the sale price (Bin, Filatova, Koning 2016, 253-54).

However, what about external factors? These include the amenities surrounding a location that buyers consider when deciding whether an area meets their needs. For example, parents may prefer homes near schools so their children can walk to them, individuals without personal vehicles might seek homes close to transportation hubs (e.g., buses, trains, or subways), and others may prioritize proximity to parks or employment centers. Research has shown that such amenities can positively influence housing prices (Esri 2019; Orford 2000). Conversely, certain externalities—such as proximity to industrial areas or airports—can reduce property values (Kiel and McClain 1995, 247-53).

With the advent of big data and machine learning, there is now an opportunity to more precisely quantify the impact of POIs on residential property valuation. Now that machine learning algorithms offer a promising alternative for real estate appraisal, this would allow real estate developers to construct properties in more accessible and valuable locations. Traditionally, ordinary least squares (OLS) regression has been the standard approach; however, OLS models are highly sensitive to outliers and often suffer from severe multicollinearity (Tomal 2020, 346). As a result, alternative methods, such as Gradient Boosting or Random Forest, should be considered, as they can predict property prices with greater accuracy than traditional hedonic pricing models (Antipov and Pokryshevskaya 2012; Flanagan and Peterson 2009).

**Objective: What is the Convenience Factor?**

A geospatial analyst examines the qualities of a neighborhood and quantifies them, measuring environmental influences as complementary to a property’s internal features. For instance, satellite imagery can be used to predict housing prices. Stephen Law et al. demonstrated this concept by showing that neighborhood characteristics were more predictive of housing prices than local streets or the houses themselves (Law, Paige, Russell 2019, 16). However, the question remains: which POI-derived features are the most predictive of residential property values?

Whatever the amenity, one consistently stands out as an essential service: the convenience factor—grocery stores and hybrid shopping centers that provide essential goods and services necessary for modern-day living. Without these modern grocery centers, the convenience factor would be lost, severely hampering everyday functioning. At best, homeowners would need to shop at local farms; at worst, they would have to travel miles to obtain essential supplies. The suggestion, then, is that the distance to the nearest grocery store is a significant factor in determining housing prices.

Therefore, we aim to determine how grocery stores influence current listing prices, similar to other external amenities. Since these amenities can be both categorized and quantified, they allow for the implementation of machine learning models. The next section explains the rationale behind the categories chosen for our models and provides support for the so-called convenience factor. We then outline the basic research methodology, which incorporates a custom-built HTML app utilizing the Overpass API. The study is designed to address the following questions:

1. Which POIs are most predictive of current listing prices?
2. Do traditional or non-traditional grocery stores have a greater impact on predicting housing costs?
3. Do grocery stores influence housing prices relative to a baseline comparison?

Once the data is analyzed using various machine learning methods—such as the suggested Random Forest—we discuss further optimization and evaluate the potential advantages and limitations of these approaches.

**Literature Review**

**Categorization**

It has already been established that external amenities can have some influence on housing prices. This effect can be calculated using hedonic pricing models, machine learning models, and geospatial analysis. But how exactly should grocery stores be classified and quantified? Ephraim Leibtag (2005) outlined a foundational framework for categorizing “grocery stores.” He divided them into **traditional grocery stores**, such as conventional supermarkets, specialty stores, and warehouse stores, which tend to be limited in size. In contrast, **non-traditional grocery stores** include supercenters, wholesale clubs, and other mass merchandisers, which are not only large but also hybrid in the variety of goods they sell.

Forrest Stegelin (2016) proposed a similar classification, focusing specifically on North Georgia and the Atlanta Metro counties. He categorized grocery stores into regional chains, warehouse clubs, national chains, mass merchants, and department stores. However, these classifications largely align with those outlined by Leibtag (2005). For instance, warehouse clubs correspond to wholesale clubs, as exemplified by Costco and Sam’s Club, while mass merchants refer to supercenters, like Walmart and Target.

Emerson et al. (Emerson, House, Palma 2003) and Joseph Brown (2004) proposed a different type of categorization, dividing grocery stores into three groups based on quality, price, and services provided. These categories include: Type A stores (higher price, higher quality), Type B stores (medium quality, medium price), and Type C stores (low quality, low price). However, this classification does not align well with the frameworks proposed by Leibtag (2005) and Stegelin (2016), as it lacks a “traditional” versus “non-traditional” distinction. Instead, the primary argument for this approach is that it simplifies the identification of correlations between customer shopping patterns, allowing one to discern potential income brackets based on store Type.

We adopted a combined approach, incorporating both a traditional versus non-traditional classification and a division based on quality and services. Palma et al. (2003) compares these types of stores to Stegelin’s (2016) mass merchants, such as Walmart, which aligns with Leibtag’s (2005) classification of supercenters. Therefore, our more eclectic approach still aims to align with the framework established by the existing literature.

**Grocery Stores as a Determinant**

Homeowners typically choose a grocery store based on its utility, which is defined by its convenience (distance from the home) and its assortment. The definition of assortment varies across studies. For instance, Brown (2004) considers it in terms of variety, quality, and price, while Mohsin Shahzad et al. (Shahzad, Zafar, and Zulqarnain 2015) argue that it is based on prices, product quality, and branding. While all these factors influence store choice, location was found to be a statistically significant factor in consumer decision-making, with more than half of respondents preferring to shop within their residential area. Respondents who favored supercenters and wholesale stores were only moderately influenced by location (Shahzad, Zafar, Zulqarnain 2015, 1170–71).

Interestingly, Brown argued that store loyalty might occasionally outweigh distance as a factor, but he also suggested that distance can create an “inertia,” causing individuals—particularly those with low income—to default to closer stores (Brown 2004, 2–3). Palma et al. (2003, 9) add that while family income can influence the choice of grocery store type, “education, gender, and age appear to have little systematic effect on store choice.” This potentially reduces concerns about other factors skewing the results.

In line with Brown’s (2004) conclusion that price influences store choice, Leibtag (2005) made a similar claim, further arguing that non-traditional stores consistently offered lower prices than traditional ones. Therefore, it can be expected that lower-income families would be more likely to choose non-traditional stores. However, Stegelin (2016, 81–82) presented mixed results on this issue. He found that supercenters, like Walmart, received lower scores compared to other grocery chains in the Southeast region, including those with comparable prices, with regional supermarkets ranking highest. Additionally, he noted that shoppers, regardless of income, tended to be "promiscuous" with their spending.

Peter Batt and Norshamliza Chamhuri (2011) also confirm that the most influential factors in consumer choice of retail food are proximity, price, and food quality. They further support Brown’s (2004) point about the importance of income, stating that demographic characteristics matter, and reiterate that higher-income earners are more likely to choose locations based on convenience. Lower-income groups, on the other hand, tend to value "traditional retail formats." Ultimately, they argue that "consumers will decide where to shop based on the minimum travel time to the nearest retail store" (Batt and Chamhuri 2011, 3). Modern retail formats in central locations attract higher-income earners who have access to cars, while the "traditional" retail format is preferred by lower-income individuals due to its closer proximity and lower associated travel costs.

A study conducted in the Northeast region on supercenters (Target and Walmart) sought to determine if these conclusions held true. Supercenters are typically classified as nontraditional shopping locations that offer lower prices, but Stegelin (2016) argued that price did not significantly impact consumer choice compared to other chains. According to the results, when the total distance was 800 meters, there was an average decrease in housing sale price, with the largest decrease occurring in the 400-599.9 meter range. However, all other distance ranges (0-199.9, 200-399.9, and 600-800 meters) experienced an increase in sale price. It is important to note that all supercenters were located in economically disadvantaged neighborhoods (Caceres and Geoghegan 2017).

**Conclusion: Proximity or Density?**

The literature generally supports the assertion that the distance from grocery stores affects sale prices. Additionally, the type of grocery store appears to play a significant role, as consumers may either be unable to afford certain stores, prefer a variety of selections, favor specific brands, or may be forced to choose a type of store based on its distance from their home.

Most of the research has focused on proximity, typically classified as "proximity to the closest POI," as the key factor. It appears that low-income individuals are often compelled to rely on distance, but the data remains inconclusive regarding which type of store influences this decision. Interestingly, development trends seem to drive rental price increases. Even in economically disadvantaged neighborhoods, the development of new grocery stores is associated with a positive rise in rental prices, particularly for homes closest to the new developments.

Therefore, we propose adding another feature in addition to proximity: density. Retailers typically locate their stores in central areas, so it can be expected that these central locations—or areas with a high concentration of grocery stores—should positively influence housing prices, possibly more so than distance alone. This is because density better reflects urbanization and development. In highly urbanized regions, distance alone may be misleading, as the same distance could encompass multiple stores within a short radius around the home. We aim to compare density with proximity to determine which produces a stronger model and which offers more significant predictive value.

**Methodology**

**Research Design**

Drawing from various influences in the literature, we based our research design around the primary target: appraisal prices of houses. The southeastern U.S., specifically Atlanta, Georgia, emerged as a "hotspot" in several studies, so we chose the Atlanta Metropolitan Statistical Area as the source for our housing data. The dataset includes typical features found in regional reports, such as sale price, square footage, age, number of bedrooms, and geolocation coordinates.

Continuing with the combined approach, we created a mixture of categories for the grocery store "features," dividing them into non-traditional and traditional types (Stegelin 2016, 82). The traditional features are as follows:

1. Supermarkets: A full service grocery stores that often sells a variety of non-food products as well. These are almost always part of a chain (e.g., Publix, Harris Teeter, Piggly Wiggly, BiLo, Ingles, Bells, Earthfare).
2. Variety Stores: Retailers that sell inexpensive items, typically with a single price point for all products (e.g., Dollar Tree, Family Dollar, Dollar General).

The non-traditional features include:

1. Warehouse Club: Non-traditional membership-retailer hybrids that sell bulk products in a warehouse environment, with at least 40% of products devoted to groceries (e.g., Costco, Sam’s Club, BJ’s).
2. Supercenter: Large food-and-drug store combinations that also sell mass merchandise. At least 40% of the products are devoted to groceries, and the store must be at least 40,000 square feet (e.g., Target, Walmart).
3. Convenience Store: Non-traditional, limited stores that sell a variety of general merchandise, including packaged food products (e.g., 7-Eleven, Quick Trip).

Before we could build the models, we first needed to collect data for the distance-based and density features. Following previous research, we defined distance as the "nearest POI." However, defining density was more challenging due to the inconclusive results in Caceres and Geoghegan (2017). To address this, we decided to define density in three ways: within a one-mile, three-mile, and a five-mile radius around the desired location. With this framework in place, we were then able to assert H0: that neither distance nor density has an effect on housing prices. H1 would be that distance from the nearest grocery store has a significant influence on housing prices. H2 is that the number of grocery stores has a significant influence on housing prices.

We encountered an immediate challenge in collecting the data for the proposed features. Typically, programs like OpenStreetMap or the Google Places API would be used for this task. However, the requirement was to ensure that both property and POI data were geocoded for spatial analysis and that spatial joins could be performed to link properties with nearby POIs. Standard APIs do not offer this functionality. Therefore, we developed an app using HTML that utilized Overpass to extract both distance and density values based on POI type, or, in some cases, “brand name.”

Once all the data was collected, missing values, outliers, and inconsistencies in the datasets needed to be addressed before the main design could be proposed. Since Antipov and Pokryshevskaya (2012) demonstrated that a Random Forest model is suitable for this type of design, we chose it as our initial model. A Random Forest model combines multiple decision trees, averaging the predictions from subsets of data to make final predictions. It is capable of capturing non-linear relationships and interactions between amenities and housing prices. However, a common issue with Random Forests is that they can struggle with imbalanced data, leading to bias toward certain features, while still incurring a high computational cost.

Therefore, we opted to use an XGBoost model as a supplement. XGBoost is a type of gradient-boosting model that builds decision trees sequentially, with each iteration improving upon the last by accounting for errors. XGBoost is faster, more powerful, and better at handling larger datasets, including those with missing values, which is common with housing data. Since we aimed to not only assess feature importance but also analyze correlations, we incorporated regression and regularization techniques alongside the tree-based methods.

Multiple linear regression is the most basic form of regression, but it cannot be tuned, so we also used Lasso and Ridge regression models. A Lasso model shrinks the coefficients to nearly zero, while Ridge regression shrinks the coefficients toward zero, but never to exactly zero. These methods help us understand the linear relationship between housing prices, external amenities, and other features.

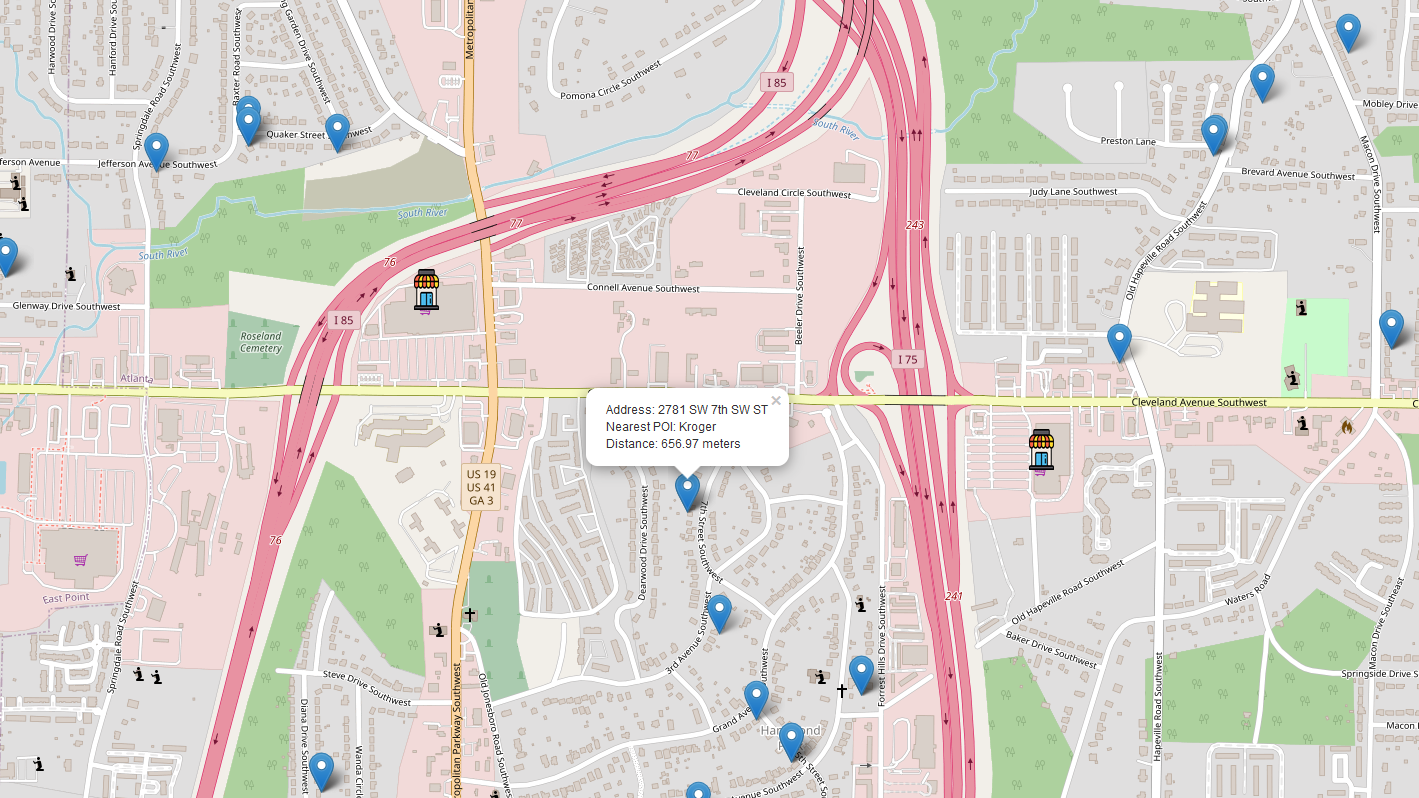
Once the models were selected, we used the XGBoost and Random Forest models to determine feature importance, as well as partial dependency, which shows the relationship between a selected feature and the target while averaging out the effects of all other features. The regularization techniques allowed us to assess the correlation, or the strength of the relationship, as well as feature strength, based on a Lasso regularization display.

**Using Leaflet: Data Collection**

To gather the density and distance values, an app was created using HTML as the base. It relies on Leaflet, an open-source JavaScript library for interactive maps, and utilizes the OpenStreetMap API, a geographic database available to anyone and updated by the community. Since geospatial data was being used, the original housing data had to be converted into a GeoJSON format so it could be read by Leaflet. In HTML, boundaries were set for the Atlanta area.

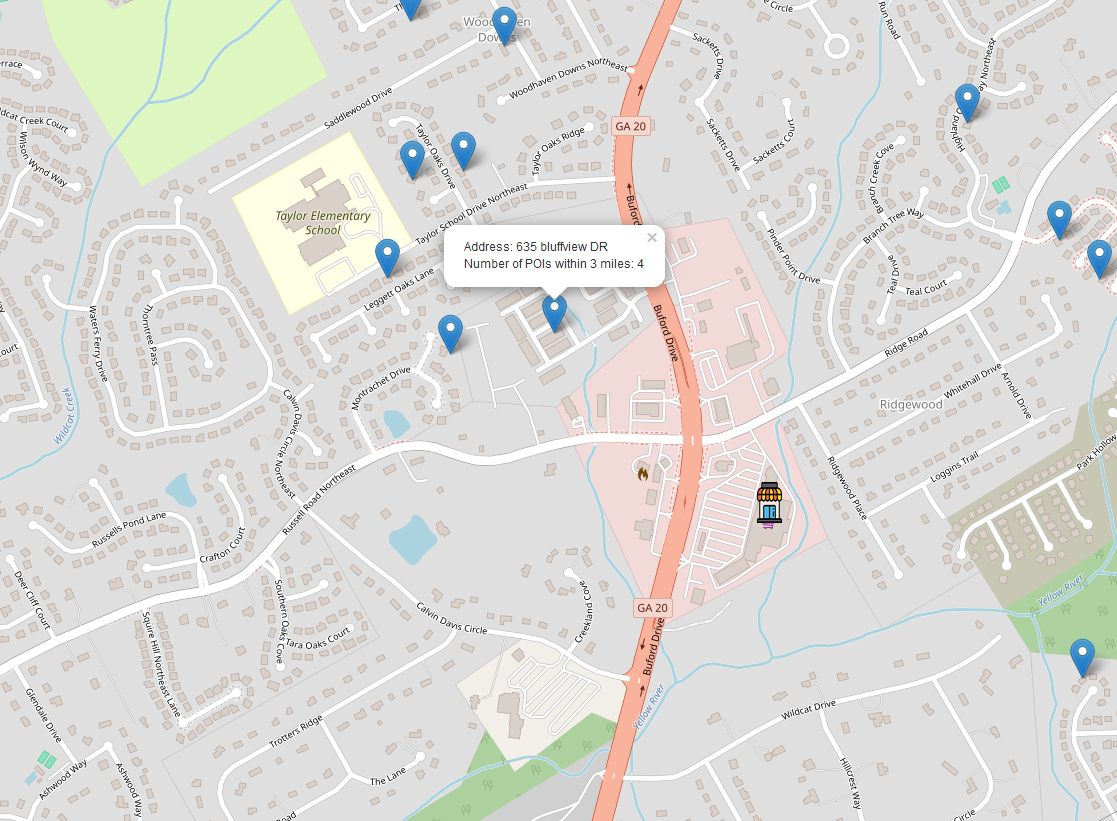
Calculating distances was fairly straightforward and done in one function. Assuming each POI had an address, the closest POI was reported using a function already built into HTML. The results were then recorded into a CSV file for later use. Overpass uses a tag-based system where POIs, such as gas stations or parks, are classified under categories based on their features. Alternatively, it can classify POIs by brand name if the store doesn’t fit a specific category. We based our traditional and non-traditional features to work well with Overpass’s tag-based system.

Figure 1 shows the distance-based map for a specific address near Cleveland Ave SW, Atlanta, GA. When a user clicks on a blue location marker, the app displays the address, the name of the nearest POI (in this case, a Kroger supermarket), and the distance to that POI. The tags we used included 'shop=supermarket', 'shop=variety\_store', 'shop=wholesale', 'shop=convenience', as well as brand names for stores that could be classified as supercenters, such as Target, Walmart, and, technically, Amazon and Aldi. However, there weren’t enough of these stores in the region to be considered relevant.



(Fig. 1 shows a distance-based Overpass search, which would be downloaded into a CSV file and used as data. Note that there is a 'shopping cart' icon at East Point that does not have the grocery store marker. This is because we were searching for supermarkets, and the store at that location is a Walmart, classified as a supercenter. Certain brands can be excluded from Overpass searches, or entirely excluded in the code.)

The density-based code functioned similarly, but instead of focusing on the closest POI, it counted the number of POIs within a defined radius (converted from one, three, or five miles). After defining the radius, the app functioned in the same way. Figure 2 shows the density-based map for an address near Taylor Elementary School and the Buford Expressway. We note that it is near a Kroger, which the distance-based version would identify and display the distance to. However, since this is the density-based model with a three-mile radius, it shows the number of supermarkets within that range, most of which are outside the current zoom level.



(Fig. 2 of density-based Overpass search of what would be downloaded to a CSV and used as data. One may note that a Kroger is closest, but only the number of POIs are shown and reported.)

Once this process is completed for the distance and the one-mile, three-mile, and five-mile radii, the data can then be combined with the original dataset based on the type of POIs that were searched. Specifically, there are now twenty features based on the five we outlined, each grouped by whether they are distance-based or density-based. If the data is density-based, it must be split between the three different radii.

**Data Analysis: Scores**

Does distance and/or density impact residential property values? Since distance was already hinted at having an effect on property values, there are two reasons for establishing H2. Proximity to certain amenities can influence property desirability. Additionally, different radii based on density can capture both local and broader neighborhood effects. A cross-tabulation was used for all models to test the R², mean squared error, root mean squared error, and mean absolute percentage error. R² is a goodness-of-fit test, with a higher value indicating that the model explains a larger portion of the target's variability. We want that score to be higher than the baseline, which is a model without any added features.

Mean squared error (MSE) and root mean squared error (RMSE) are loss metrics that calculate the squared difference between predicted values and target values, averaged over all observations. RMSE simply takes the square root of the result, making it more interpretable by expressing the error in the same unit as the target values. A smaller value indicates better model performance, and, unlike R², these scores do evaluate predictive accuracy. Mean absolute percentage error (MAPE) measures a model's prediction accuracy relative to actual values, with the error expressed as a percentage. For example, MAPE is scale-independent, with a smaller MAPE indicating better predictive accuracy.

Once the default parameters were identified, we tuned each model using specific hyperparameters tailored to the characteristics of each model. Hyperparameter tuning allows certain model parameters to be refined in an attempt to improve the score, such as tree depth and the alpha value. We then used the 'best parameters' to obtain the desired results.

We used a train-test-validation split to account for overfitting and analyze all data. The training set is used for initial fitting but isn't needed for the final evaluation. The validation set (20% of the data) is used for tuning and assessing model performance. The combined train-validation set allows us to train on 80% of the data, capturing more patterns. The test set is the main goal. Scores derived from it indicate how well the model is likely to perform on truly unseen data. The closer the test score is to the train-validation score, the less overfitting there is.

**Distance-Based Scores**

Only one set of results is reported for the distance-based models, with each model showing varying results. Tables 1 and 2 displays the scores for the tuned baseline scores. The baseline serves as a confirmation of our null hypothesis, using only the original features before adding the distance and density-based features. If any model fails to outperform the baseline, it indicates that the new features did not improve predictive power. Worse performance would suggest that the added complexity introduced more noise rather than enhancing predictions. While the Ridge regression achieved a slightly higher R² than the Lasso, the Lasso model was chosen for consistency (see Appendix E1).

Table 1: Score Comparison for XGBoost Baseline

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4764 | 0.5805 | 0.5875 |
| **MSE** | 441042.7435 | 344791.1960 | 385699.1357 |
| **RMSE** | 664.1105 | 587.1892 | 621.0468 |
| **MAPE** | 0.1686 | 0.1561 | 0.1585 |

Table 2: Score Comparison for Lasso Baseline (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation** |
| **R2** | 0.3493 | 0.3275 | 0.3608 |
| **MSE** | 0.0720 | 0.0739 | 0.0750 |
| **RMSE** | 0.2683 | 0.2718 | 0.2739 |
| **MAPE** | 0.0237 | 0.0239 | 0.0242 |

Table 3 displays the XGBoost scores for the tuned distance model. The slight overfitting observed, indicated by the 17.01% difference between the train-validation and test sets, was intentional. We could have made the train-validation scores nearly equal to the test scores, but that would have significantly lowered the test scores. Therefore, for each model, it is expected that the train-validation scores will be higher than the test scores, and subsequently the validation scores to be higher as well.

Typically, a higher R² is desired, but it’s important to note that R² does not measure predictive power. The MAPE of 14.93% indicates that the model's predictions deviate from the actual values by an average of 14.93%, which is more informative. The RMSE alone is not particularly insightful, as its value is influenced by the scale of the target data. Therefore, it is more useful when compared to the RMSE values of other models. Luckily, this is our first solid example of a non-baseline model being superior to the baseline.

Table 3: Score Comparison for XGBoost Distance

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6042 | 0.7165 | 0.7170 |
| **MSE** | 333416.2553 | 232965.2795 | 264625.4240 |
| **RMSE** | 577.4221 | 482.6648 | 514.4176 |
| **MAPE** | 0.1493 | 0.1319 | 0.1334 |

In comparison to this model, the Random Forest iteration of the distance-based data performed worse than the XGBoost model in every metric (see Appendix A2). The R² was lower, while the MSE, RMSE, and MAPE were higher than their XGBoost counterparts. This pattern was consistent across all subsequent density-based variations, confirming that the Random Forest model may not be the most suitable approach for this dataset.

Table 4 shows the multiple linear regression (MLR) scores for the distance-based data. The MLR is an alternative approach to the XGBoost and Random Forest models, with different parameters to tune. Linear models are generally used for smaller datasets and assume linearity. In this case, the target was log-transformed to improve performance. While the scores may appear superior to those of the XGBoost model, they are not. If the XGBoost model had also been log-transformed, it would have outperformed the MLR in every metric, particularly when comparing the test set scores against those of the Lasso (see Appendix Group D).

Table 4: Score Comparison for MLR Distance (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3498 | 0.3394 | 0.3928 |
| **MSE** | 0.0717 | 0.0723 | 0.0711 |
| **RMSE** | 0.2662 | 0.2682 | 0.2656 |
| **MAPE** | 0.0235 | 0.0234 | 0.0234 |

The regularization models both showed improved scores compared to the linear regression model, but their results were very similar to each other. This can be attributed to the fact that linear models are not designed to capture non-linear relationships or handle the complexity that XGBoost models can. Table 5 shows the Lasso regression scores for the distance-based data, which, notably, only had a higher R² than the Ridge model (see Appendix E2). Despite the improvements from both the Lasso and Ridge models, they did not offer a significant enough boost in performance to overcome the distance-based metrics, with the exception of the MAPE.

Table 5: Score Comparison for Lasso Distance (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3778 | 0.3570 | 0.3881 |
| **MSE** | 0.0688 | 0.0706 | 0.0718 |
| **RMSE** | 0.2624 | 0.2658 | 0.2680 |
| **MAPE** | 0.0230 | 0.0233 | 0.0236 |

**Density-Based Scores**

The density-based scores were divided into three sets based on the radius. The one-mile radius had the fewest POIs, simply by default, and it also had the highest number of locations with zero POIs within the selected radius, which can significantly impact the results. Table 6 demonstrates just how low the scores can be, even lower than the distance model. Interestingly, despite having an inferior score compared to the XGBoost distance model, the log-transformed scores, including the MAPE, are actually superior to all distance-based linear models. As in the previous section, the Random Forest model performs worse in every metric compared to the XGBoost model, though its scores are not low enough to be worse than the distance-based linear models (see Appendix A3).

Table 6: Score Comparison for XGBoost One Mile

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5394 | 0.6371 | 0.6300 |
| **MSE** | 387973.6284 | 298231.6637 | 345940.0183 |
| **RMSE** | 622.8753 | 546.1059 | 588.1667 |
| **MAPE** | 0.1613 | 0.1478 | 0.1507 |

Table 7 shows that, among all the linear models, once again the Lasso is the superior one. However, this is more evident here (see Appendix E3). The Lasso model has a higher R² than the Ridge model and lower RMSE and MSE values. Unlike the XGBoost model, the one-mile linear model outperforms the distance-based model in every metric. This trend is not limited to the Lasso model; each linear model for the density-based data shows higher scores in all categories compared to the distance models.

Table 7: Score Comparison for Lasso One Mile (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3977 | 0.3693 | 0.3954 |
| **MSE** | 0.0666 | 0.0693 | 0.0710 |
| **RMSE** | 0.2581 | 0.2632 | 0.2664 |
| **MAPE** | 0.0228 | 0.0232 | 0.0236 |

Our first alternative hypothesis, H1, posits that housing prices are influenced by their proximity to grocery stores. The results currently show that the distance model outperforms the one-mile density model. However, H2 suggests that as the number of grocery stores in proximity increases, housing prices will also rise. Therefore, as we expand the radius, we would expect the stores to become more relevant as features.

The three-mile radius highlights the impact that increasing the radius can have on the results. Tables 8 and 9 show a clear improvement in all scores compared to the one-mile results as the radius expands. As expected, the Random Forest model still produced inferior scores overall, although it did show improvement relative to the one-mile results. Interestingly, it also managed to slightly surpass the distance-based XGBoost scores. The three-mile density-based Random Forest model demonstrated a much larger improvement compared to the distance-based Random Forest model (see Appendix A4). For the Lasso model, only its R² was superior to the Ridge model (see Appendix E4).

Table 8: Score Comparison for XGBoost Three Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6065 | 0.7330 | 0.7355 |
| **MSE** | 331476.7945 | 219402.3056 | 247331.0408 |
| **RMSE** | 575.7402 | 468.4040 | 497.3239 |
| **MAPE** | 0.1451 | 0.1281 | 0.1292 |

Table 9: Score Comparison for Lasso Three Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4566 | 0.4341 | 0.4672 |
| **MSE** | 0.0601 | 0.0622 | 0.0625 |
| **RMSE** | 0.2452 | 0.2493 | 0.2501 |
| **MAPE** | 0.0216 | 0.0221 | 0.0222 |

Table 10 shows a continued increase in scores, though the improvement is less pronounced compared to the jump from one mile to three miles. At this radius, the scores now clearly surpass the XGBoost distance scores. Notably, for the second time, the validation set achieves a higher score than the train-validation set, despite this model not exhibiting the highest level of overfitting. Table 11 further demonstrates a plateau in improvements, as the scores show only minimal gains over the three-mile model.

Table 10: Score Comparison for XGBoost Five Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6307 | 0.7382 | 0.7497 |
| **MSE** | 311078.9243 | 215139.6725 | 234024.0497 |
| **RMSE** | 557.7445 | 463.8315 | 483.7603 |
| **MAPE** | 0.1417 | 0.1265 | 0.1287 |

Table 11: Score Comparison for Lasso Five Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Test Set** | **Train-Val Set** | **Validation** |
| **R2** | 0.4682 | 0.4571 | 0.4893 |
| **MSE** | 0.0588 | 0.0596 | 0.0599 |
| **RMSE** | 0.2425 | 0.2442 | 0.2448 |
| **MAPE** | 0.0211 | 0.0216 | 0.0219 |

**Score Evaluation: Baseline**

Overall, the distance-based models achieved higher scores than the density-based models up to the three-mile radius. Notably, at no point did the distance-based linear models outperform their density-based counterparts. As expected, the density-based scores improved as the radius increased, capturing more POIs within the specified range. The XGBoost five-mile model emerged as the best-performing model among the twenty-five tested, achieving the highest scores across all metrics except for MAPE, assuming a log transformation was applied. However, the five-mile linear model compensated for this by consistently producing the lowest MAPE values across all models.

Because the distance-based models outperformed the one-mile models, we can attempt to draw a few conclusions. One possibility is that distance plays a more significant role, initially, when fewer grocery stores are considered. It is important to note that many houses had no grocery stores within one mile (for certain categories), which inherently gives the distance-based models an advantage. Another explanation is that the comparison between the two approaches may not be entirely fair, as they could be measuring similar information. Most grocery stores of a given category are likely located within one to two miles of a home. As a result, both the distance-based data and the one-mile data might effectively capture the same grocery store, just presented in a slightly different manner.

Regardless, the baseline had the worst overall scores in every category when compared to all other models. From these results, we can conclude that predictive power improves as the number of POIs increases and that proximity plays an important role. However, can we safely reject the null hypothesis based on these findings? The answer is likely no. While the results demonstrate differences between the models, they do not necessarily indicate whether these differences are statistically significant or meaningful.

Table 12 displays the results of a paired-samples t-test, comparing the baseline model to the distance-based model to determine if the two are statistically different. The results show that every metric fell below the 0.05 threshold, indicating a statistically significant difference. Additionally, Cohen’s d, which measures substantive significance, demonstrated at least a small effect size for each metric.

Table 12: Significance Between Distance and the Baseline XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 12.98040 | 0.0 | 4.10476 |
| **MSE** | 12.91056 | 0.0 | 4.08268 |
| **MAPE** | 24.89549 | 0.0 | 7.87265 |
| **RMSE** | 14.50156 | 0.0 | 4.58580 |

The same findings apply to the multiple linear regression for the distance-based model, which showed strikingly similar results despite being a different type of model, as seen in Table 13. This pattern is repeated when comparing the density-based models to the baseline (Tables 14–16). Notably, the one-mile density model showed the least improvement over the baseline, which aligns with the previously observed score trends.

Table 13: Significance Between Distance and the Baseline Lasso Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 11.27056 | 0.0 | 3.56406 |
| **MSE** | 11.65826 | 0.0 | 3.68667 |
| **MAPE** | 10.98819 | 0.0 | 3.47477 |
| **RMSE** | 13.42285 | 0.0 | 4.24468 |

Table 14: Significance Between Density One Mile and the Baseline XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 10.14706 | 0.0 | 3.20878 |
| **MSE** | 10.63525 | 0.0 | 3.36316 |
| **MAPE** | 15.71720 | 0.0 | 4.97021 |
| **RMSE** | 10.95962 | 0.0 | 3.46573 |

Table 15: Significance Between Density Three Miles and the Baseline XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 14.88335 | 0.0 | 4.70653 |
| **MSE** | 12.73996 | 0.0 | 4.02873 |
| **MAPE** | 32.45928 | 0.0 | 10.26453 |
| **RMSE** | 15.07872 | 0.0 | 4.76831 |

Table 16: Significance Between Density Five Miles and the Baseline XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 20.24031 | 0.0 | 6.40055 |
| **MSE** | 12.79712 | 0.0 | 4.04680 |
| **MAPE** | 38.37219 | 0.0 | 12.13435 |
| **RMSE** | 17.27271 | 0.0 | 5.46211 |

**Score Evaluation: Model Comparison**

While it has already been established that the baseline is the weakest model and that distance and proximity are relevant, the focus now shifts to comparing the models against each other. Specifically, the goal is to determine whether the distance-based model differs significantly from the density-based models and, subsequently, whether the density-based models are significantly different from one another. Table 17 presents results that align with expectations: the mean values for the distance-based model are consistently higher than those for the one-mile density model across all metrics.

Table 17: Significance Between Distance and Density One Mile XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-value** | **Cohen's d** |
| **R²** | 11.79338 | 0.0 | 3.72939 |
| **MSE** | 11.36362 | 0.0 | 3.59349 |
| **MAPE** | 17.54096 | 0.0 | 5.54694 |
| **RMSE** | 12.73930 | 0.0 | 4.02852 |

While the score comparison initially indicated that the distance-based model outperformed the one-mile density model, the density-three model marginally exceeded the distance-based model in terms of scores. Table 18 reveals an interesting reversal: although Tables 1 and 6 showed close scores, with Table 6 being superior, the t-statistic in Table 18 shifts in favor of the density-three model. The results are statistically significant, contrary to initial expectations. However, as anticipated, the effect size remains very small. Table 19 further illustrates this trend, showing that the differences between models grow as the density radius increases. Notably, the difference between density-one and the distance was by far the strongest.

Table 18: Significance Between Distance and Density Three Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -4.90789 | 0.00084 | -1.55201 |
| **MSE** | -4.08738 | 0.00273 | -1.29254 |
| **MAPE** | -5.84076 | 0.00025 | -1.84701 |
| **RMSE** | -4.44782 | 0.00161 | -1.40652 |

Table 19: Significance Between Distance and Density Five Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -7.53910 | 0.00004 | -2.38407 |
| **MSE** | -5.16756 | 0.00059 | -1.63413 |
| **MAPE** | -9.85704 | 0.00000 | -3.11707 |
| **RMSE** | -6.08206 | 0.00018 | -1.92332 |

We also sought to determine if there is a statistically significant increase between each successive density model. Tables 20 and 21 show that, as the radius increases, the statistical significance actually diminishes, with the p-value approaching the 0.05 threshold. Additionally, the effect size weakens, and the differences between the model values, as predicted, begin to stagnate. This trend persists despite the density-five model achieving the highest scores across all models.

Table 20: Significance Between Density One and Density Three Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -15.15527 | 0.0 | -4.79252 |
| **MSE** | -11.05147 | 0.0 | -3.49478 |
| **MAPE** | -22.69907 | 0.0 | -7.17808 |
| **RMSE** | -13.69943 | 0.0 | -4.33214 |

Table 21: Significance Between Density Three and Density Five Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -5.70063 | 0.00029 | -1.80270 |
| **MSE** | -4.63896 | 0.00122 | -1.46697 |
| **MAPE** | -8.45679 | 0.00001 | -2.67427 |
| **RMSE** | -5.08650 | 0.00066 | -1.60849 |

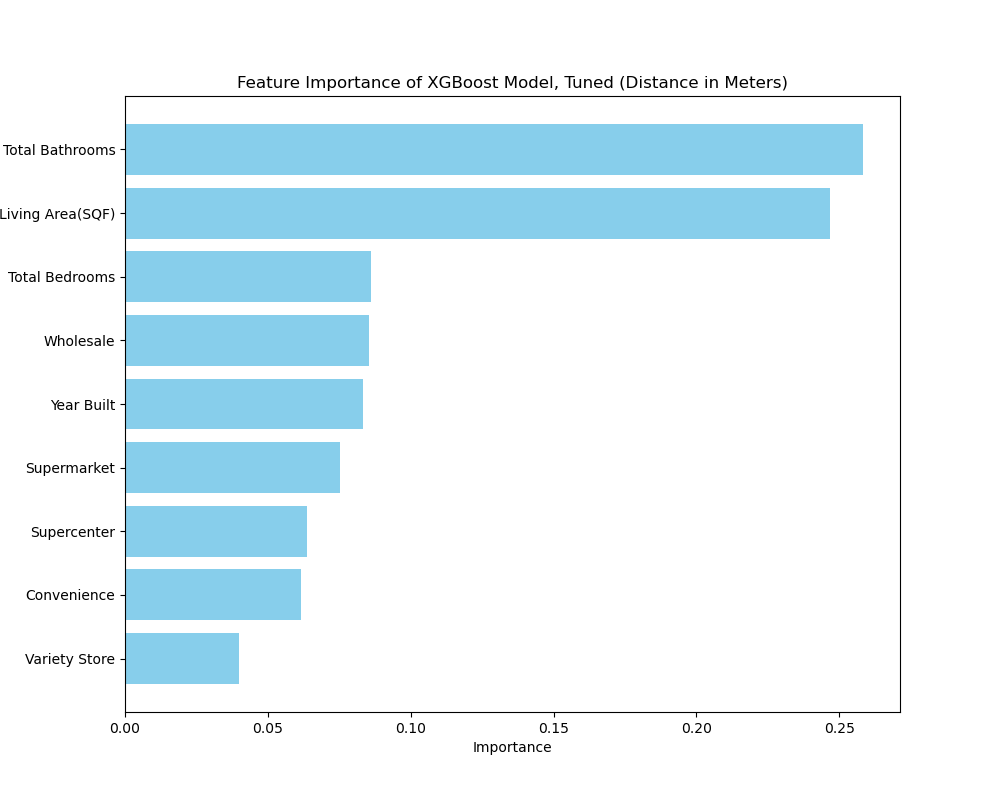
**Data Analysis: Feature Impact**

**Feature Importance**

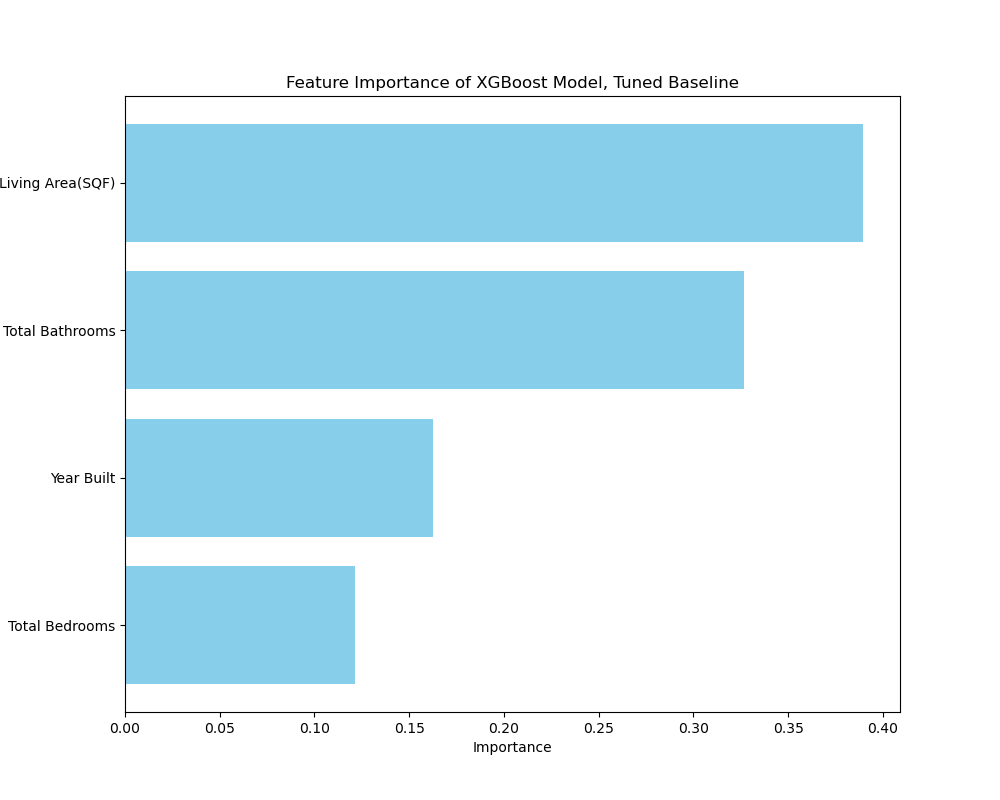
Having determined which models are significant and which outperform others, the next objective is to assess which features influence housing prices. Feature importance refers to the extent to which each feature contributes to the model's predictions. The Random Forest and XGBoost models compute feature importance in different ways. XGBoost emphasizes improvement over folds, whereas Random Forest focuses on the reduction of impurity through fold splits. Despite these differences, both models can be used for feature ranking, and a higher importance indicates that a feature has a more substantial impact on performance.

Figure 3 displays the feature importance for the XGBoost distance model. It reveals that Total Bathrooms, Living Area, and Total Bedrooms rank highest in importance. Notably, these baseline features outperform the distance features in terms of feature importance. The first distance feature to appear is 'Wholesale,' followed by the final baseline feature, and then the remaining distance features. 'Variety Store' exhibits low feature importance, indicating that it contributes minimally to the model's predictive power.

Figure 4 displays the feature importance for the baseline model, which closely resembles the layout of the distance feature importance. The key differences are that 'Year Built' now surpasses 'Total Bedrooms,' and overall importance has increased. This is not surprising, as the two features originally had similar values, and multicollinearity is now less of a concern. It is important to note that in Figure 4, 'Living Area' is now the most significant predictor, surpassing 'Total Bathrooms.' However, in the density models (Figures 5 and 6), 'Living Area' holds greater importance. In the Random Forest models (see Appendix B1&B2), 'Living Area' consistently ranks as the most important feature.

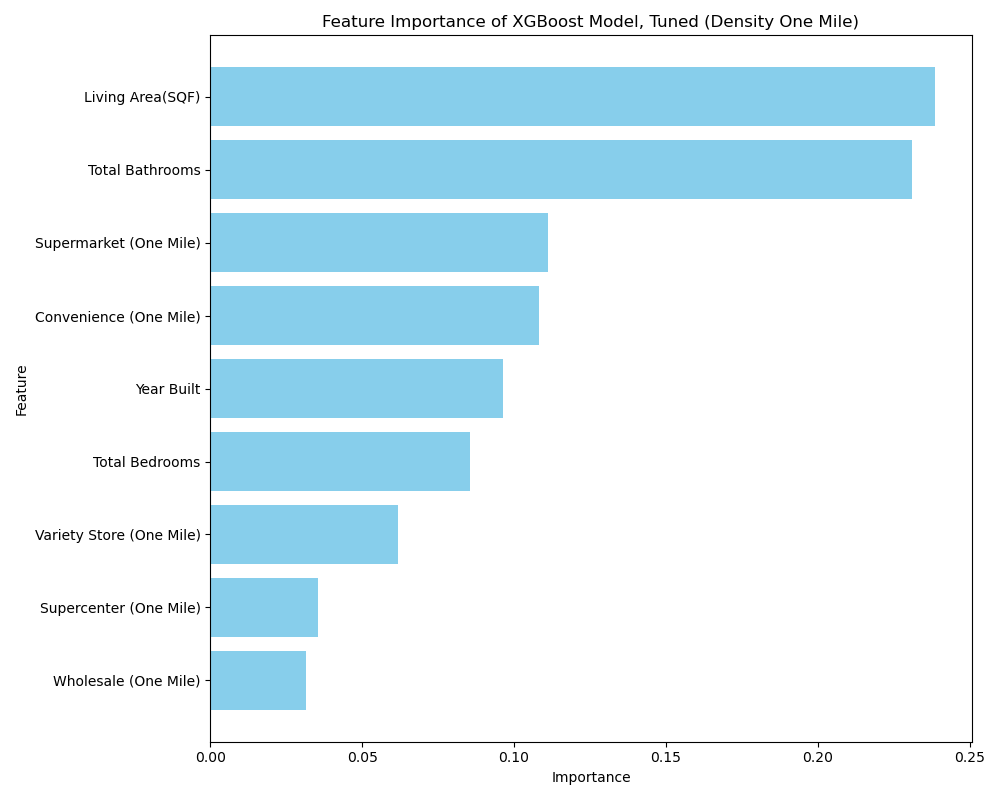


(Fig. 3: Distance XGBoost Model Feature Importance)

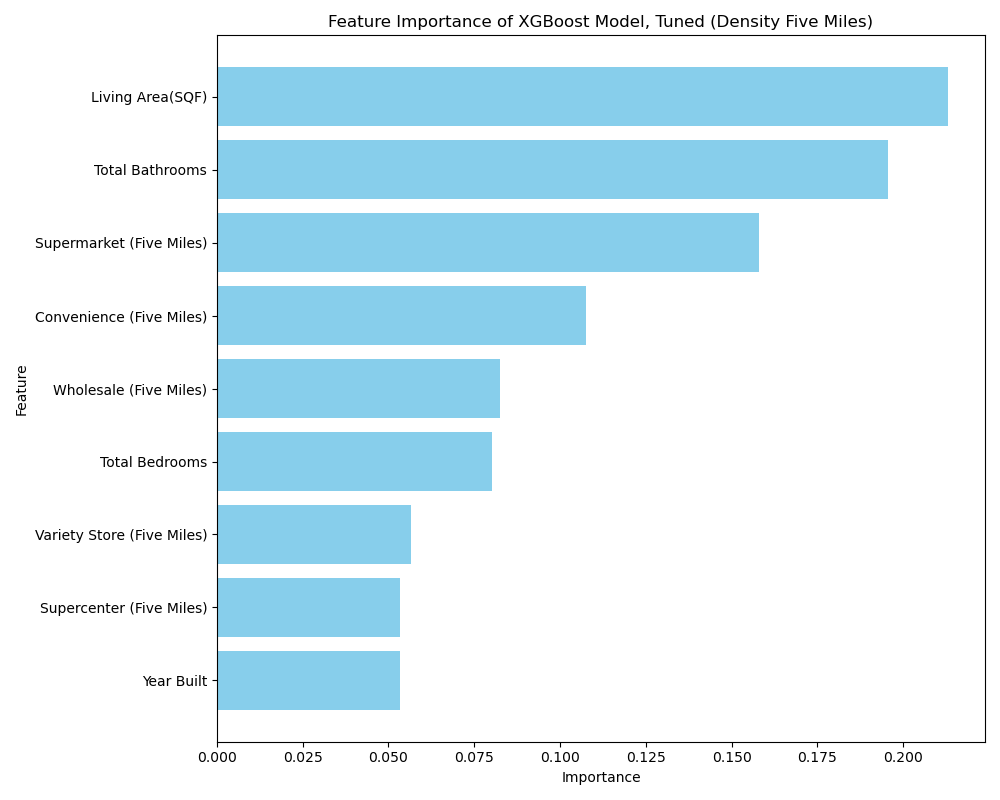


(Fig. 4: Baseline Model Feature Importance)

The density models provide clearer insights into the composition of the data. Figure 5 illustrates that supercenters and wholesale stores not only have the least predictive power, but, according to the Random Forest version (see Appendix B3), wholesale stores have no influence at the one-mile radius. This can be explained by the fact that there are too few wholesale stores to significantly impact predictive power. More importantly, we observe that supermarkets and convenience stores have managed to surpass baseline features even at the one-mile radius. In Figure 6, this trend becomes more pronounced, with the importance of the supermarket feature increasing rapidly as the radius expands. Notably, wholesale stores surpass four other features at the five-mile radius, and in the Random Forest version, supermarkets are ranked as the second most important feature (see Appendix B4).



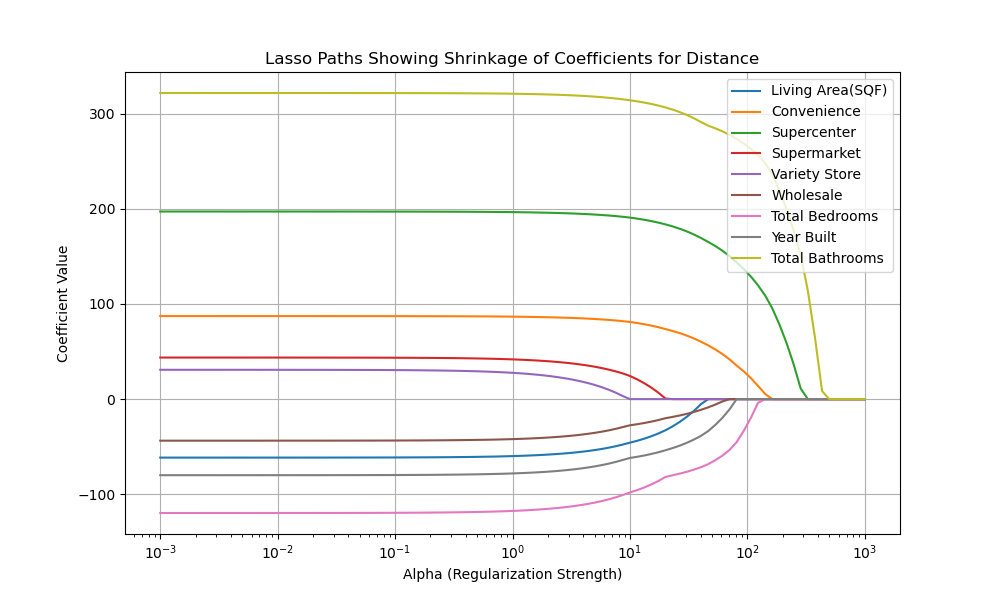
(Fig. 5: XGBoost Density One Mile Feature Importance)



(Fig. 6: XGBoost Density Five Miles Feature Importance)

Although multiple linear regression models typically do not display feature importance, it is possible to identify which features exert a stronger influence when penalized based on the size of their coefficients. Features with weaker or redundant relationships to the target variable are penalized more rapidly, causing their coefficients to shrink toward zero at lower alpha values. Conversely, features that withstand the penalty longer typically exhibit a stronger influence. It is important to note that the linear regression models yielded much lower scores than models such as XGBoost and Random Forest, so a difference in feature importance is to be expected.

This difference is evident in the distance model, as shown in Figure 7, where there is a complete reversal compared to Figure 4. For instance, while both models start with "total bathrooms" as the most important feature, the Lasso model diverges significantly in how it ranks other features. The Lasso regression applies a penalty to the coefficients and assumes no regularization, focusing only on direct linear effects. In contrast, the XGBoost model accounts for both linear and non-linear effects, which makes some features appear more important than in Lasso. Additionally, XGBoost considers feature interactions and is not dependent on scaling, unlike Lasso. While the Appendix Group F shows the density model comparisons, further comparison is unnecessary as the results differ significantly between the two approaches.



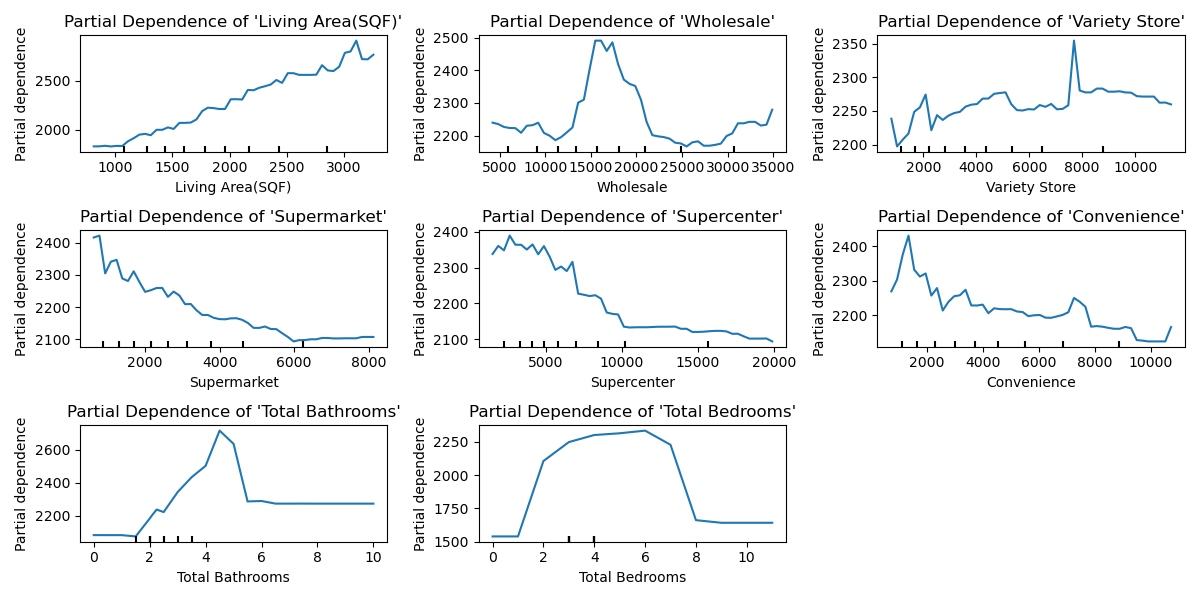
(Fig. 7: Lasso Regularization Graph for Distance Features)

**Partial Dependency**

While feature importance provides a general sense of predictive power, it offers limited insight into the specific impact of each feature on the target variable. To better understand the influence of individual features, we need to examine how predicted outcomes change as feature values vary. A Partial Dependence Plot (PDP) is a tool that isolates features to assess their effect on a model’s output. By observing how changes in a feature's value influence the predicted outcome, we can determine if the relationship is linear, exponential, or characterized by specific thresholds. This approach will help explain, for example, why the wholesale feature has such low importance.

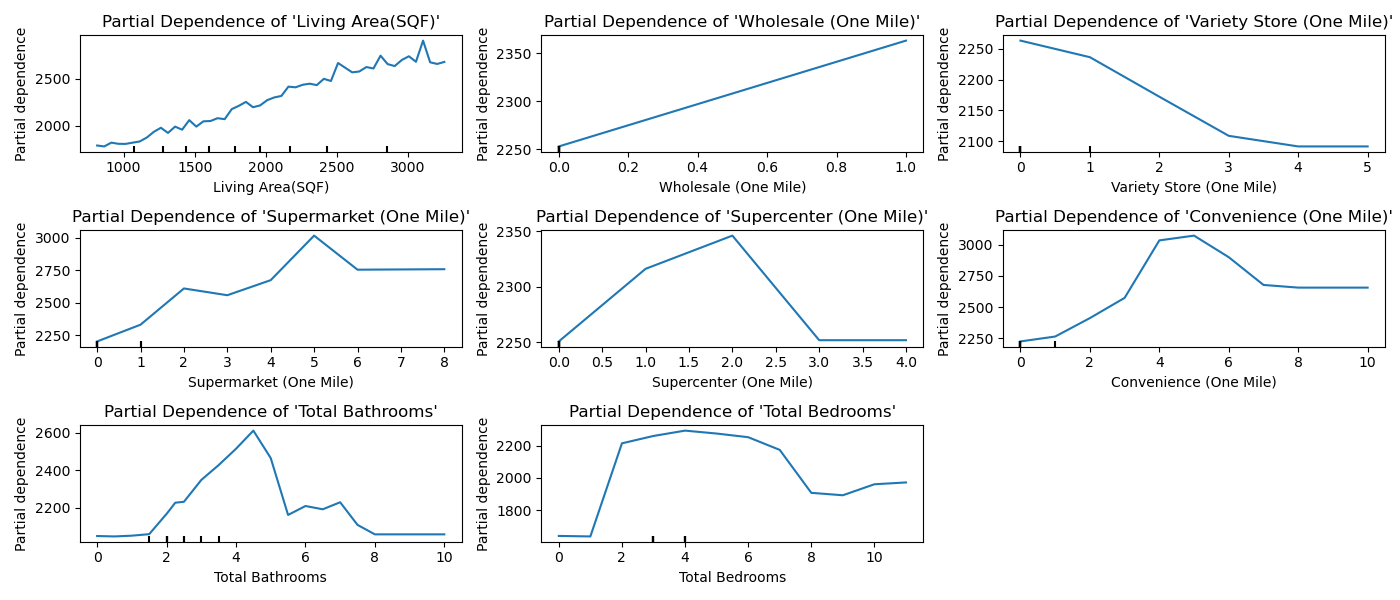
PDPs display the effect of features independently, so a separate baseline "plot" is not required. Figure 8 presents the PDP for both the baseline and distance features, using the XGBoost model and the train-validation set. In these plots, the Y-axis represents the predicted outcome, influenced by the features shown on the X-axis—specifically, the average housing prices. This visualization allows us to observe how predictions change as the values of the features vary. For instance, as mentioned in the introduction, an increase in living area corresponds to a rising trend in predicted housing prices. This can now be seen more directly in Figure 8.

More importantly, three of the distance features—supermarkets, supercenters, and convenience stores—exhibit a downward trend. This indicates that, at least initially, housing prices have higher predicted values the closer they are to these specific types of stores. As the distance to the nearest POI increases, the predicted housing price decreases. However, this pattern does not apply to wholesale stores, which seem to reach a peak around 9 miles (15,000 meters) before declining. This suggests that wholesale stores are sparse in general. In contrast, variety stores present an anomaly, as they do not follow any discernible trend.



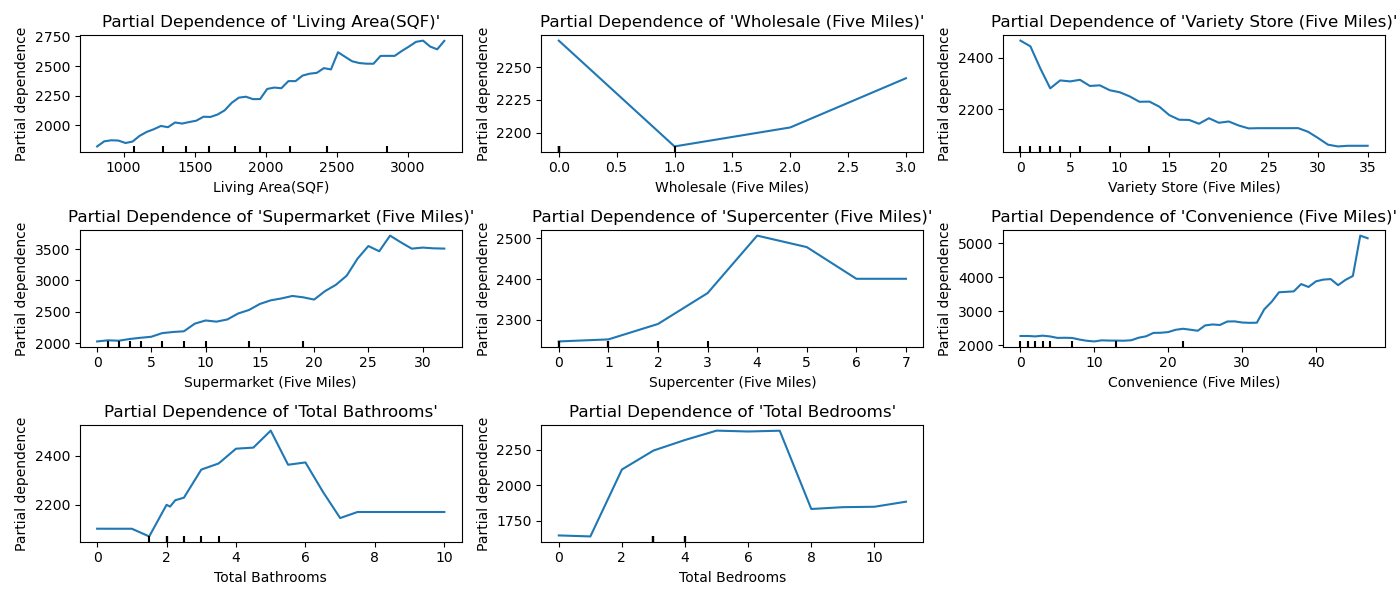
(Fig. 8: XGBoost PDP of Distance Features)

The density PDPs yield similar results to the distance PDPs, but with an opposite trend, as a higher number of stores within a given mile radius corresponds to higher values. For example, in Figure 9, we observe that supermarkets exhibit a steady increase in predicted housing prices. In other words, the greater the number of supermarkets within one mile, the more likely housing prices are to be higher. This trend holds for all non-baseline features, except for variety stores, which, as previously noted, is an anomaly. Supercenters do show a flattening trend, which may be due to the limited number of houses with four or more supercenters within the one-mile radius, preventing a discernible effect. Additionally, it is worth noting that no house had more than one wholesale store within one mile.



(Fig. 9: XGBoost PDP of Density One Mile Features)

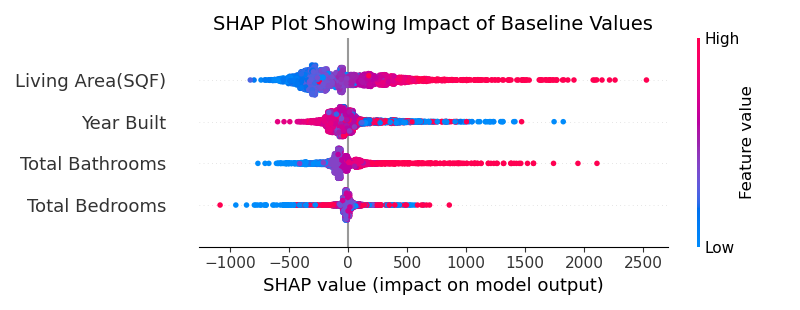
According to Table 14, the density-one model had the weakest strength against the baseline. Given that the distance scores surpassed the density-one model in the score comparison, it may be a less reliable predictor of housing prices. Fortunately, Figure 10 presents a more concise version of Figure 9, clearly showing an increase in predicted housing prices as the number of POIs increases. The wholesale feature decreases at one, which is explained by Figure 9, as no feature had more than one wholesale store within one mile. Interestingly, variety stores also continue their downward trend, potentially indicating a negative externality.



(Fig. 10: XGBoost PDP of Density Five Miles Features)

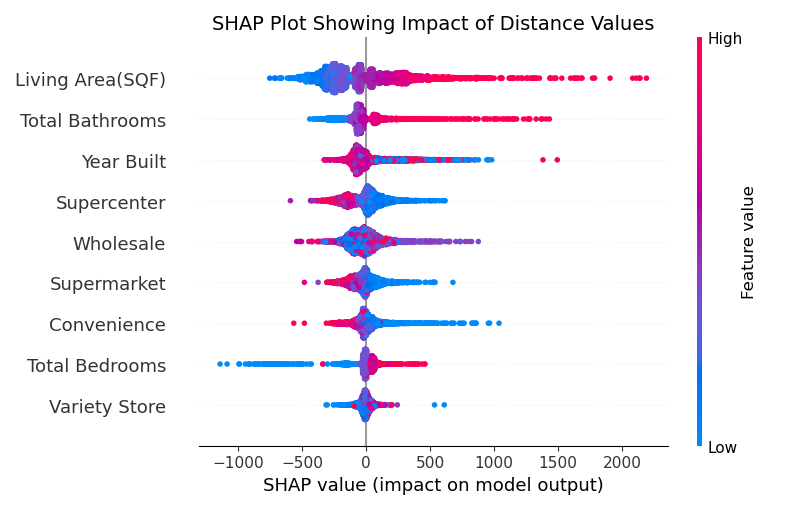
**Impact**

Now that we understand how each feature affects the predicted outcome, we can also examine the overall impact of these features using the SHAP (SHapley Additive exPlanations) plot. Essentially, we aim to determine if a feature consistently changes the target value in a specific direction for each observation. Features with positive values increase the target, while those with negative values decrease it. Figure 11 presents the baseline SHAP summary plot, which explains the overall predictions based on the data the model "sees." Each feature is interpreted differently for the baseline. For instance, in the case of Living Area, the plot indicates that "red" values (housing properties with larger living areas) tend to have a positive SHAP value, meaning they contribute positively to the predicted housing value. SHAP measures feature importance differently, ranking features based on their average impact on predictions, rather than the frequency with which the feature appears across tree splits, as is the case with XGBoost.



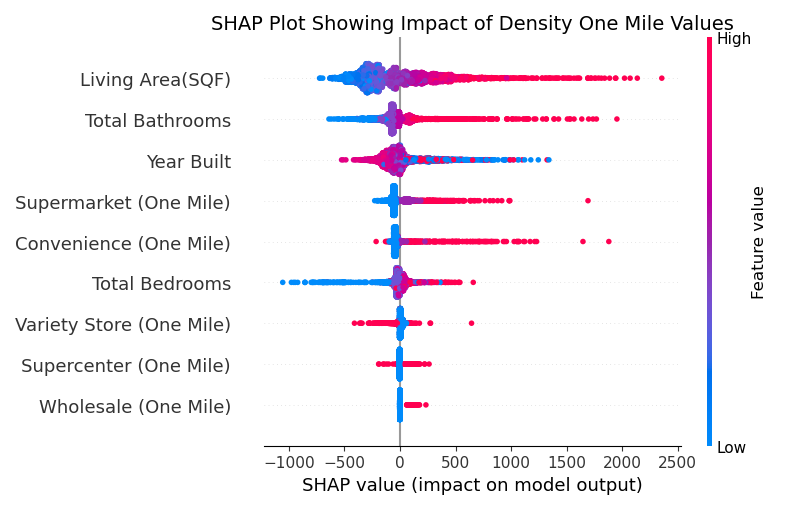
(Fig 11: SHAP Summary Plot of Baseline Features)

When considering distance, the baseline features do not appear to be significantly affected by multicollinearity. However, we can observe the trend suggested by the PDPs. For at least supercenters, supermarkets, and convenience stores, the closer the store is to a property, the greater the increase in housing value. Conversely, the opposite trend seems to hold true for supercenters, where stores located farther away appear to negatively impact predicted values. Wholesale and variety stores, however, exhibit mixed results, which may be attributed to their low feature importance and erratic partial dependencies.



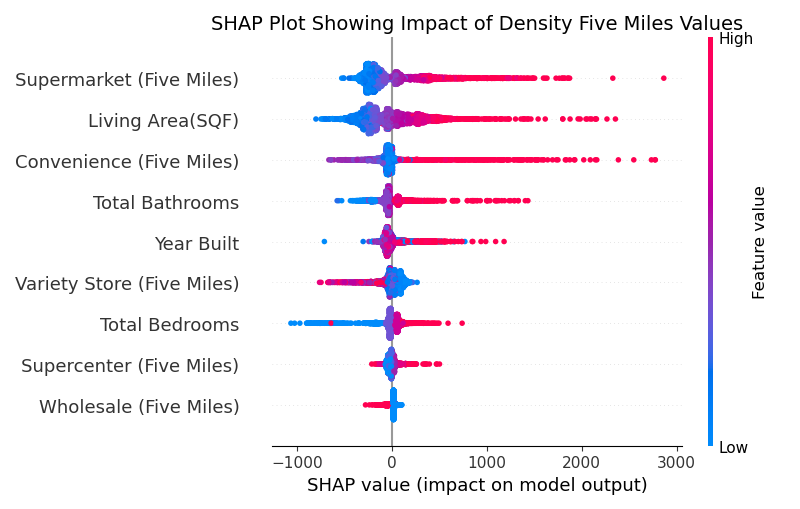
(Fig. 12: SHAP Summary Plot of Distance Features)

Figure 13 again shows a reversal in trend for the features, with a "red" high value now indicating a positive impact. It is evident that, even at one mile, supermarkets and convenience stores contribute to higher predicted housing values as their numbers increase. However, the results for the remaining store types are inconclusive, although variety stores has a slight trend towards a negative impact.



(Fig. 13: SHAP Summary Plot of Density One Mile Features)

When examining the five-mile plot, the impact becomes more pronounced. Figure 14 shows that supermarkets have emerged as the most influential feature, surpassing even living area. Convenience stores continue the same trend, while supercenters begin to exhibit a slight positive impact on housing values. However, for the second time, as shown in Figure 10, the number of variety stores appears to have a negative impact on housing values. Although a definitive conclusion cannot be made, preliminary results suggest that the number of wholesale stores also negatively affects the target.



(Fig. 14: SHAP Summary Plot of Density Five Miles Features)

**Conclusion**

The null hypothesis (H₀) posited that the selected point-of-interest (POI) category, grocery stores, does not affect housing prices. Table 12 demonstrated a significant difference between the baseline model and the features representing the distance between the POIs and house locations. Table 3 revealed that the distance-based models outperformed the baseline scores (Table 1). Furthermore, Figures 8 and 12 illustrated the importance and impact of each distance feature on housing prices, showing that, for most features, proximity to the store is associated with an increase in predicted housing values.

Given that H₁ posited that housing prices are affected by proximity to grocery stores, it was further inferred that the number of grocery stores also plays a significant role. This led to the formulation of H₂, which suggests that the density of grocery stores also influences housing prices. Tables 14 and 15 confirmed this hypothesis, demonstrating a significant difference between the density models and the baseline model. Figures 9 and 10 indicated a general increase in the predicted housing values as the values of the features increased. This trend was more clearly illustrated in Figures 13 and 14, where supermarkets emerged as the most important feature at a five-mile radius, assuming the calculation is based on average impact rather than frequency or gain.

Overall, it appears that the null hypothesis can be rejected based on both the existing literature and our findings. The general trend indicates that the closer a grocery store is to a house, the higher the predicted housing values. Furthermore, an increased number of grocery stores in proximity to the house may also lead to higher predicted values. It would seem that supermarkets and convenience stores have a tendency to be a stronger predictive factor than the other features. The only exceptions to this trend are variety stores and potentially wholesale stores, which may actually decrease housing values as their presence increases. Supercenters seem to give mixed results, showing a negative impact for distance, but not for density.

Therefore, an ideal housing location that would maximize price would be one situated in what is commonly referred to as a "convenient" area—one that is close to supermarkets and convenience stores, with a variety of these establishments nearby. The distinction between "traditional" and "non-traditional" categories does not appear to significantly impact the results, as both types of features were drawn from different categories. Additionally, it seems that being too close to a supercenter, and especially a variety store, may actually decrease property value; however, this could be an indirect effect, possibly influenced by literature suggesting that lower-income brackets tend to prefer these types of locations.

**Model Evaluation: Advantages and Disadvantages**

Five primary models were used, split into twenty-five configurations due to the use of five different data sets. The strongest model was identified as the density-five XGBoost model, while the weakest model, excluding the baseline, was the density-one multiple linear regression (MLR) model. The XGBoost models consistently delivered the best scores, while the linear regression models consistently performed the worst. The Random Forest models typically fell in between, with performance slightly lower than the XGBoost models but still superior to the regression models. XGBoost models are non-linear and capable of capturing complex relationships between the features and target variable, whereas MLR, being a linear statistical model, assumes linearity and struggles with large datasets.

Even after attempting to improve the MLR model by incorporating regularization through Ridge and Lasso, the scores remained insufficient to reach the typical performance levels of XGBoost. The limitation of both models lies in their assumption of linearity, with coefficients derived from Maximum Likelihood Estimation (MLE), which contradicts the non-linear relationships captured by XGBoost. For instance, the Lasso coefficients for distance features assumed that every distance coefficient was negative, except for the one related to variety stores.

While the XGBoost model is powerful, it is not without its drawbacks. It can be difficult to interpret directly—requiring multiple visualizations, as was the case in this study—and is highly sensitive to tuning. For instance, overfitting can occur not due to data issues, but from slight adjustments to hyperparameters such as alpha, depth, or the number of estimators. The Random Forest model shares similar challenges with respect to its parameters. Though it is simpler, it comes at the cost of computational expense, with certain parameters significantly affecting the results and requiring considerable time to tune. Given the substantial size of the dataset in this study, XGBoost was favored, as it not only addressed the research questions effectively but also did so in a timely manner.

The feature importance graphs provided a quick and effective way to display the results. They also align with the insights that a Random Forest model typically seeks to convey. While the initial results were intriguing, they did not fully address the research question. For instance, Figure 3 indicated that wholesale was the only store type more important than a baseline feature. However, without additional context, this observation raises questions about its significance. In isolation, this might suggest that we cannot reject the null hypothesis, as the other store types appear to be irrelevant. Fortunately, the density models revealed a different story, confirming that feature importance alone required further support to draw meaningful conclusions.

The PDP graphs were among the most valuable in illustrating how features influenced the target variable, particularly because each graph was independent of the others. They provide a straightforward answer: does an increase or decrease in the feature correlate with a rise or fall in predicted average housing prices? However, a limitation of PDPs is their exclusion of feature interactions, which can lead to misleading interpretations. Additionally, since "partial dependence" represents an average value, it tends to smooth data and overlook important nuances. For instance, the variety store results in Figure 8 may resemble "noise" more than meaningful trends.

Finally, we used SHAP as an attempt to visualize the actual value scores and their impact on the target variable. One immediate issue with the SHAP display is its complexity, as many of the values are centered on zero due to the nature of our data. This made the plot appear more like a frequency distribution rather than something interpretable. However, despite the noise, SHAP remains one of the most insightful visualizations. It provides a wealth of information, and without it, we would not have been able to confidently assess how the POIs interacted with the target variable.

**Evaluation of Using Leaflet:**

Leaflet proved to be an invaluable tool for collecting and visualizing geospatial data in this study. One significant advantage of implementing it through HTML was the absolute control it provided over the entire research process. This level of control allowed for precise adjustments to parameters and definitions, aligning them with Overpass's own standards to minimize confusion. As a result, the initial five categories were clearly delineated with minimal overlap. Additionally, the application was highly customizable, enabling adjustments to ranges, values, and tags as needed. With only minor modifications to the code, the application could be adapted to generate entirely different datasets, a flexibility that facilitated the production of both the density and proximity results.

One limitation of HTML is that it functions as a markup language for web content rather than a programming language. While JavaScript was embedded into the HTML to enable map interactivity, other tools such as R are more suited for data analysis and visualization. Notably, R also supports Leaflet as a package, making it an effective alternative for handling geospatial data. Consequently, utilizing R might have been a more advantageous approach for displaying Leaflet’s data. However, using R alone would lack user-friendliness. To address this, creating a Shiny application could offer a more interactive and accessible solution while eliminating the need for meticulous maintenance of HTML files.

As discussed in the data collection section, Leaflet was selected due to project constraints and its distinct advantages. However, using Overpass for tag retrieval presents certain limitations. Since Overpass is an open-source platform, it may not capture all POIs within the area. For instance, some density values might be underreported due to factors such as obscure store names or locations not yet included in the Overpass database. Fortunately, as this study focused on housing data, the omission of a few locations is unlikely to significantly affect the observed trends. In fact, such omissions are more likely to reinforce the general patterns identified in the analysis.

**Further Research**

Atlanta, Georgia was selected for this study due to prior research conducted in the region related to POIs and housing data. It serves as a well-established site for studies, being a major economic hub in the Southeastern United States and home to a prominent public transportation system. Based on the findings of this research, it is strongly recommended to further explore other regions of the U.S. using the same or similar methodologies. Alternatively, comparing these results with those from other regions could offer valuable insights into whether the demographic and economic composition of Georgia influenced the outcomes or if the impact of grocery stores on housing prices is consistent across different regions.

It is also recommended to either expand upon or narrow the focus to specific features used in this study. The features selected were influenced by a combination of Overpass's limitations and existing designs in the literature. The way these features are defined can significantly, if not fundamentally, alter the study's design. For instance, some researchers might not consider supercenters as a viable feature, opting instead to group brands like Walmart and Target under the supermarket category. Others may classify large chains like Publix as supercenters. Additionally, some might prefer to disaggregate stores into more specific categories, incorporating regional and national chains as individual features rather than using a general "supermarket" category.

In this study, we focused on density and distance, comparing them against the baseline. However, the literature suggests that other factors also influence consumer shopping behavior. Incorporating additional features, such as store quality, price, assortment, and branding, alongside proximity and density, could provide valuable insights into how these factors interact. It would be particularly interesting to examine whether some individuals might prioritize store quality over proximity, potentially leading to unexpected increases in housing prices.

Additionally, our study examined the entire Atlanta, Georgia metropolitan area, which includes both urban and rural regions. Batt and Chamhuri (2011) suggested that there may be differences in the types of income groups attracted to stores in "central locations," with lower-income consumers often being forced to shop locally due to limited mobility. However, our analysis did not account for this disparity, as it combined both urban and rural data. In other words, the impact of grocery store proximity on housing prices may differ significantly between central urban areas and rural regions, potentially leading to varying effects on housing values.

Regarding the models used in this study, three key points stand out: the R² scores were consistently below 0.90, the reliance on a single type of model may have limited the analysis, and although some overfitting was intentionally permitted, it was still observed. Ideally, the design would feature a range of models that compete against one another, with certain models yielding similar performance metrics, rather than having all models show statistically significant differences in their results. To improve the model performance, one possible approach would be to incorporate Kernel Ridge regression as a complement to Ridge regression. Given the strong performance of XGBoost, another avenue for exploration could be Stochastic Gradient Boosting, which introduces randomness to improve model generalization. Additionally, a hybrid model, such as a Voting Regressor, could be considered, which combines predictions from multiple models through weighted averaging. Finally, rather than relying solely on XGBoost, a specialized gradient boosting method, such as LightGBM, could be explored, as it models uncertainty by estimating probabilities instead of just point estimates.

**Appendix**

**Group A: Random Forest Score Comparison**

Table A1: Score Comparison for Random Forest Baseline

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4703 | 0.5817 | 0.5952 |
| **MSE** | 446192.9639 | 343822.8303 | 378467.4515 |
| **RMSE** | 667.9768 | 586.3641 | 615.1971 |
| **MAPE** | 0.1703 | 0.1589 | 0.1605 |

Table A2: Score Comparison for Random Forest Distance

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5561 | 0.6678 | 0.6671 |
| **MSE** | 373959.0615 | 273041.3606 | 311283.0150 |
| **RMSE** | 611.5219 | 522.5336 | 557.9274 |
| **MAPE** | 0.1583 | 0.1469 | 0.1487 |

Table A3: Score Comparison for Random Forest One Mile

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5250 | 0.6104 | 0.6041 |
| **MSE** | 400140.1595 | 320230.5750 | 370136.6712 |
| **RMSE** | 632.5663 | 565.8892 | 608.3886 |
| **MAPE** | 0.1650 | 0.1561 | 0.1588 |

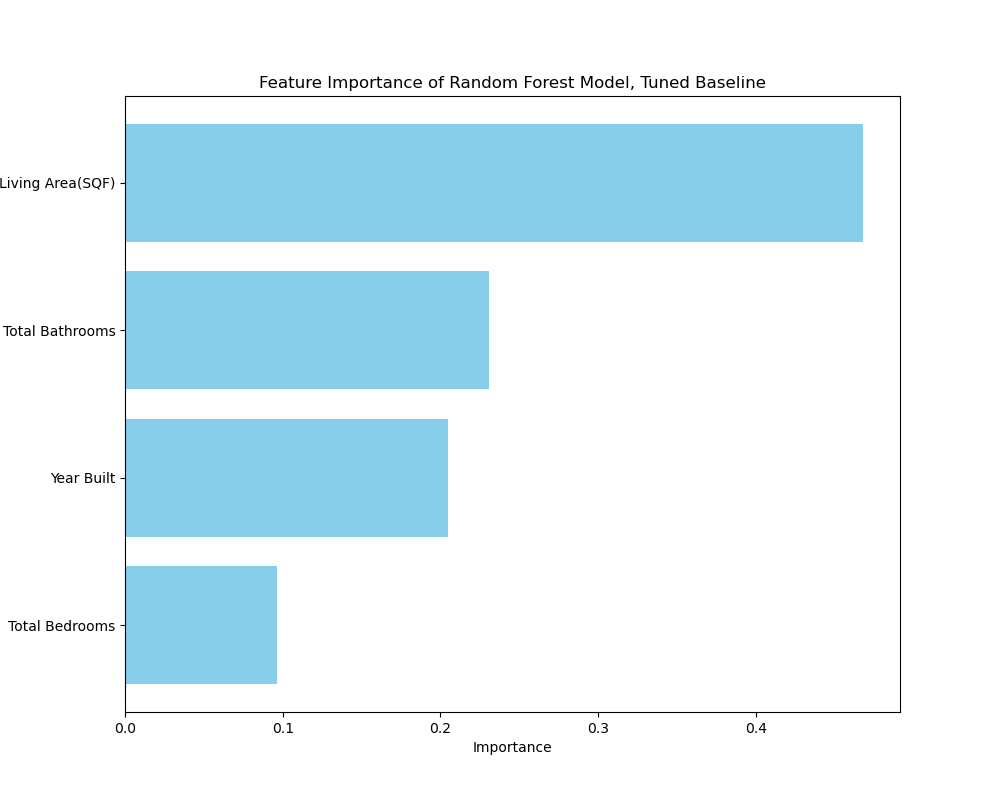
Table A4: Score Comparison for Random Forest Three Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5965 | 0.6950 | 0.5712 |
| **MSE** | 339854.8841 | 250670.3866 | 400960.4861 |
| **RMSE** | 582.9707 | 500.6699 | 633.2144 |
| **MAPE** | 0.1479 | 0.1371 | 0.1523 |

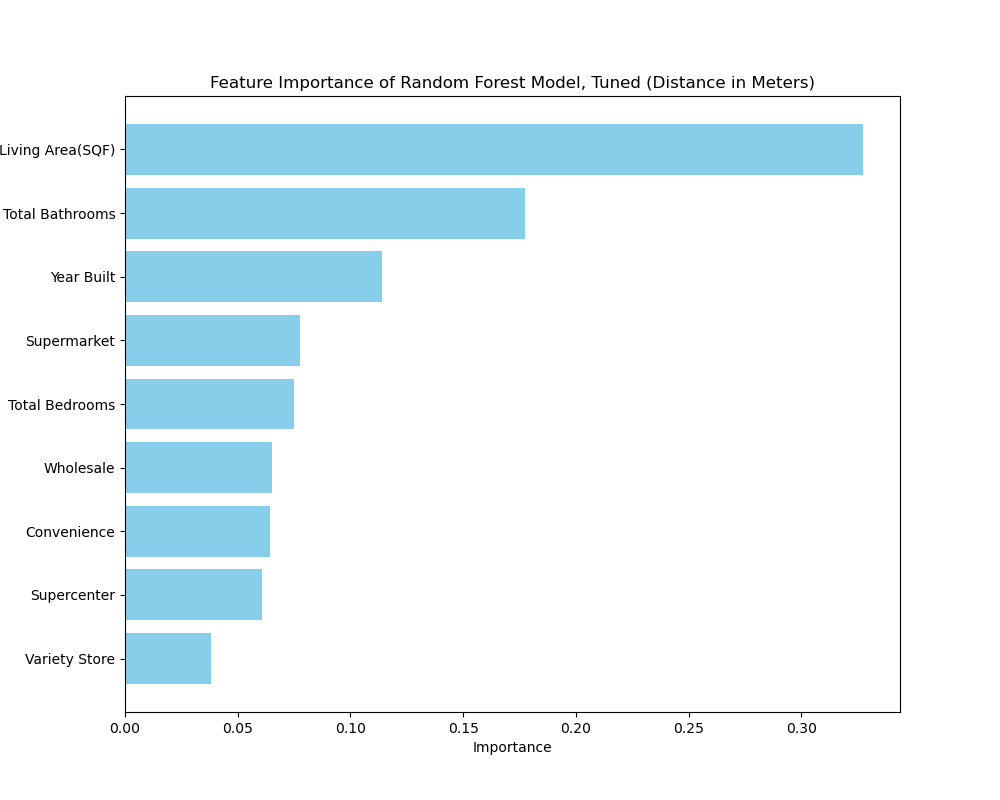
Table A5: Score Comparison for Random Forest Five Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6238 | 0.7243 | 0.7332 |
| **MSE** | 316860.6756 | 226578.5219 | 249454.5543 |
| **RMSE** | 562.9038 | 476.0026 | 499.4543 |
| **MAPE** | 0.1435 | 0.1335 | 0.1350 |

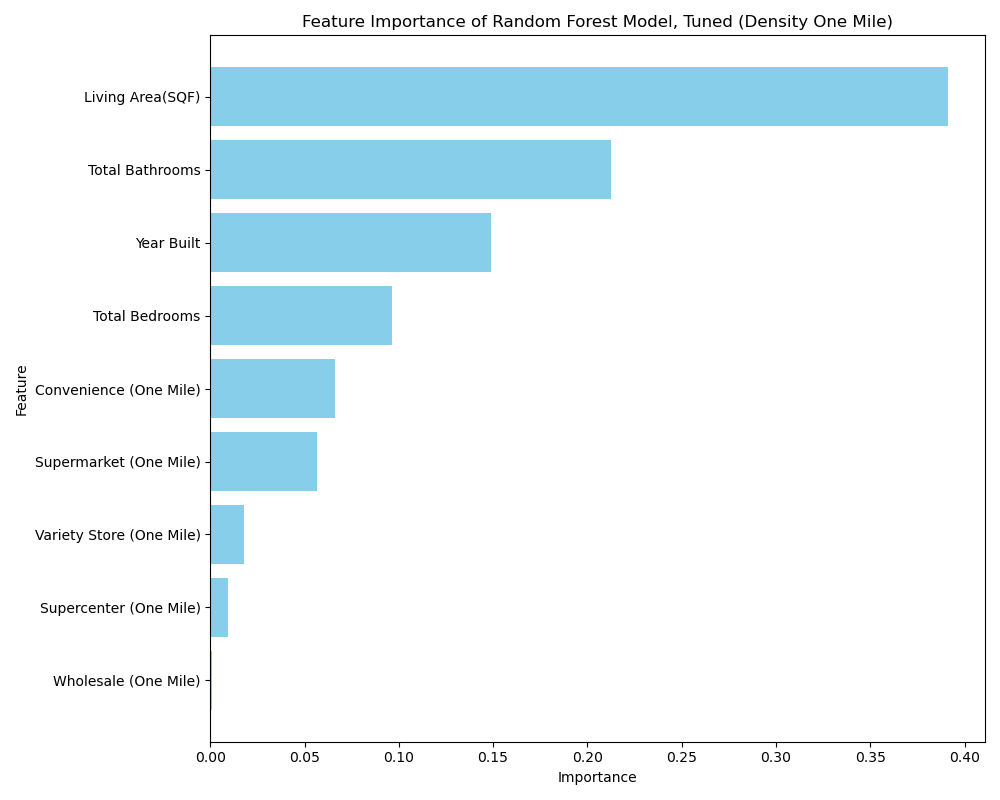
**Group B: Random Forest Feature Importance**

****

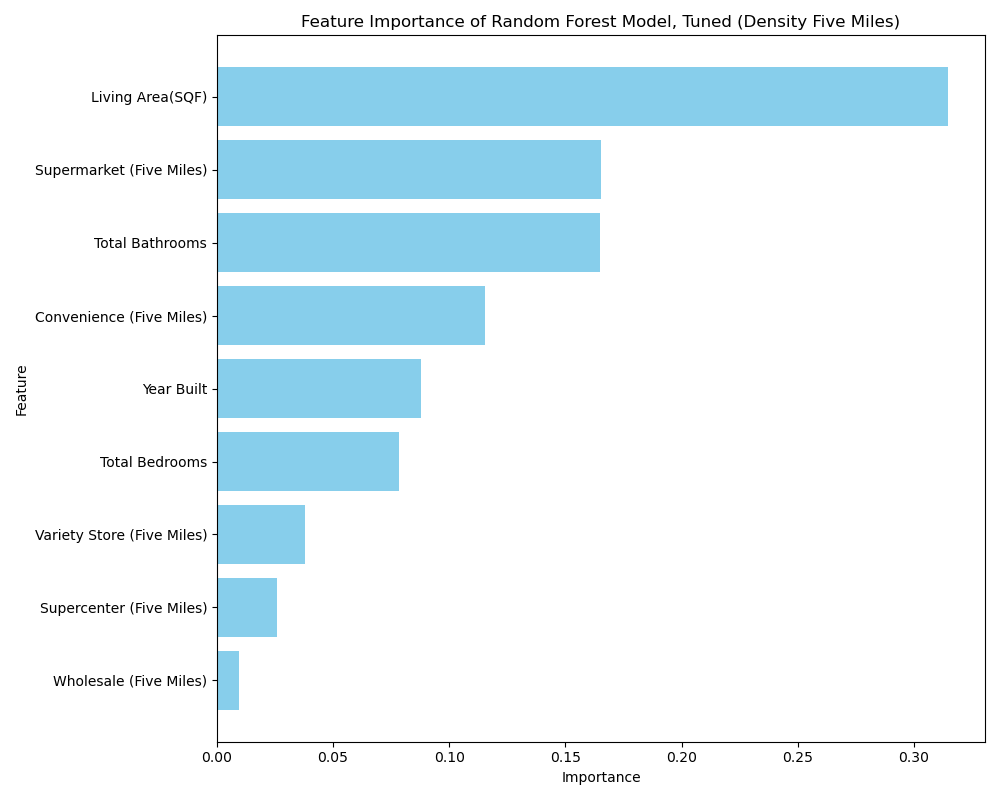
(Fig. B1: Random Forest of Baseline Feature Importance)

****

(Fig. B2: Distance Random Forest Model Feature Importance)

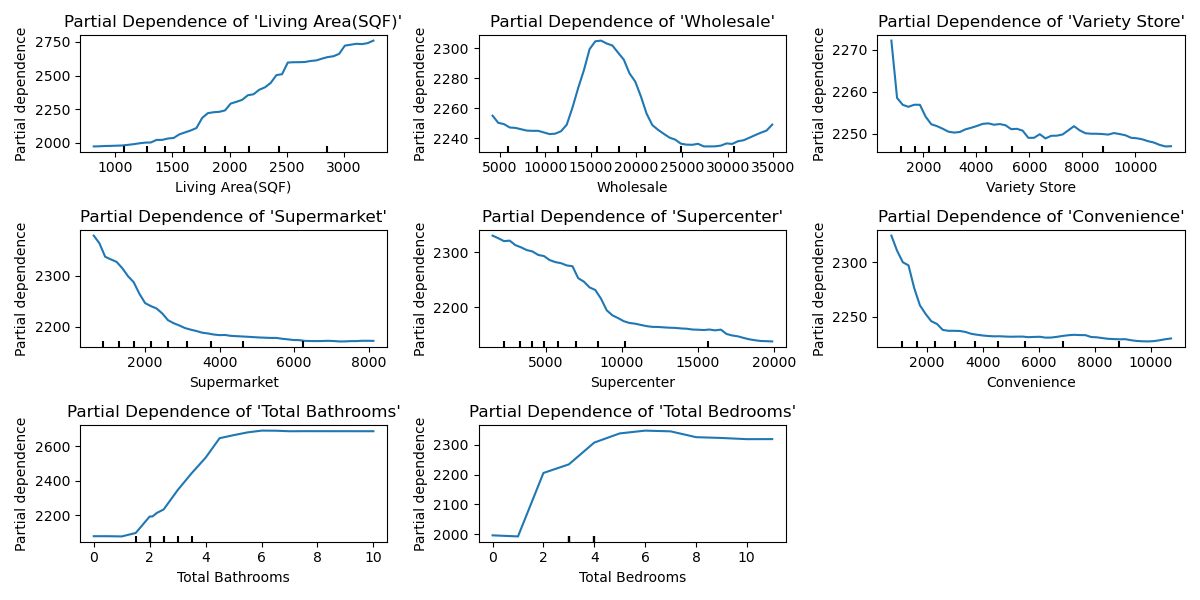


(Fig. B3: Density One Mile Random Forest Model Feature Importance)

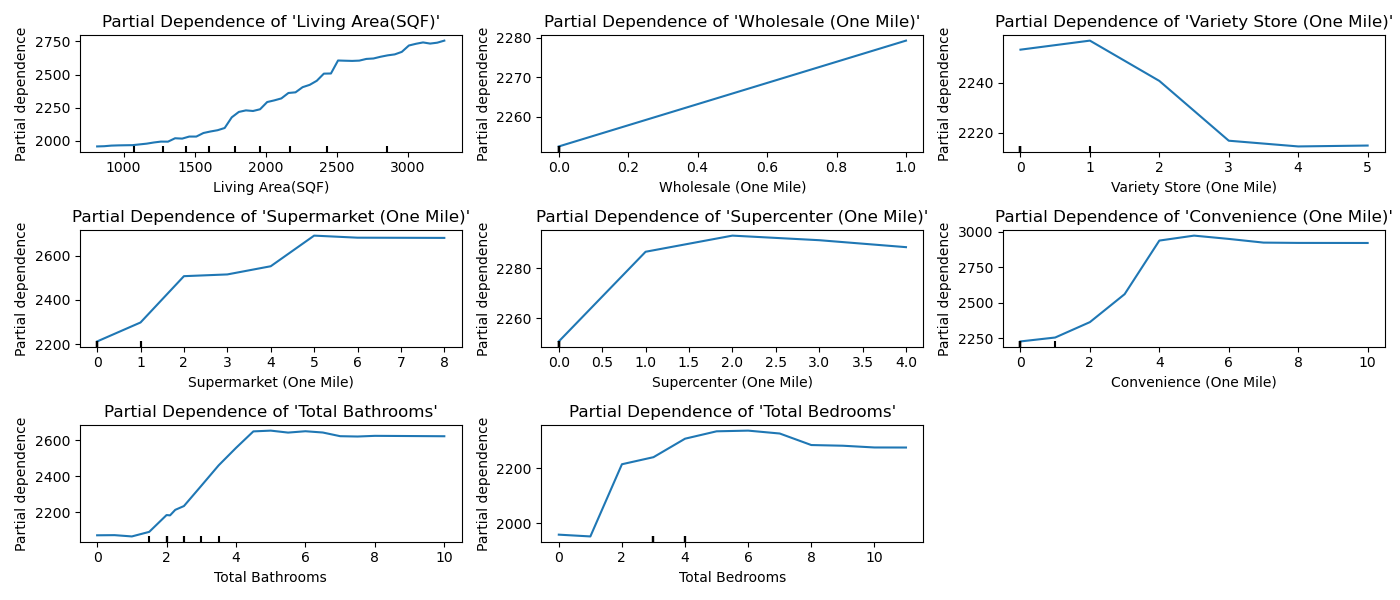


(Fig. B4: Density Five Miles Random Forest Model Feature Importance)

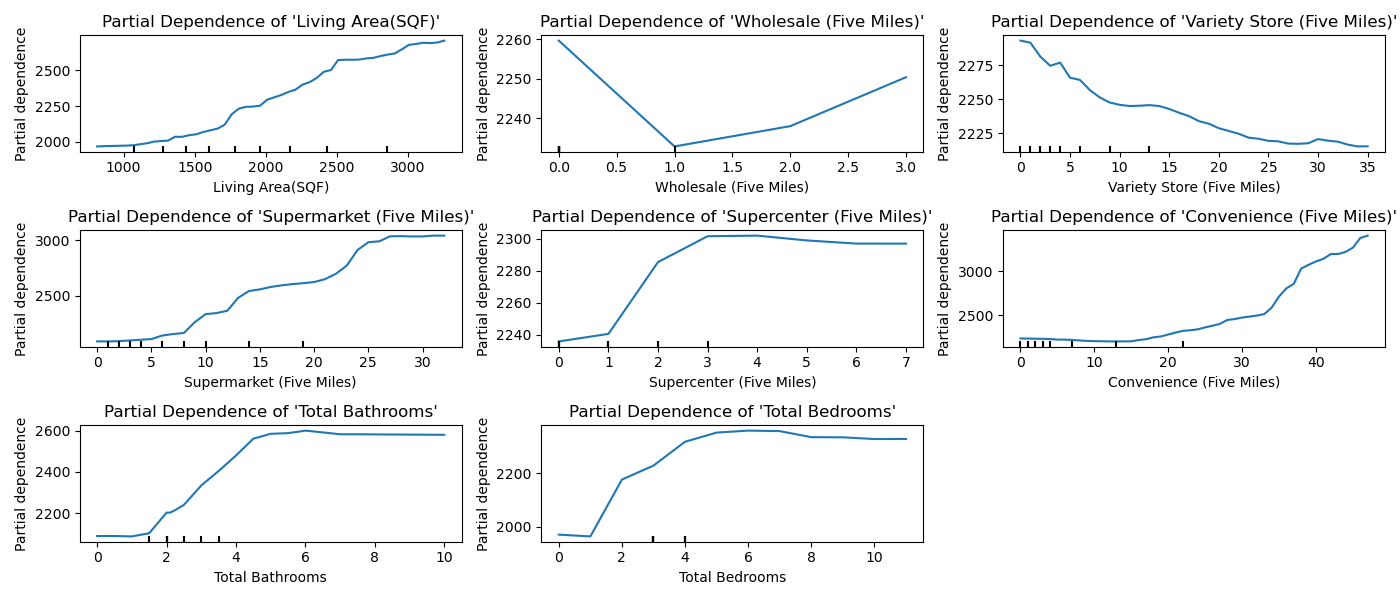
**Group C: Partial Dependence Plot**

****

(Fig. C1: Distance Random Forest Model Feature Importance)



(Fig. C2: Density One Mile Random Forest Model PDP)



(Fig. C3: Density Five Miles Random Forest Model PDP)

**Group D: Log-Transformed XGBoost Score Comparison**

Table D1: Score Comparison for XGBoost Baseline (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4920 | 0.5035 | 0.5124 |
| **MSE** | 0.0562 | 0.0545 | 0.0572 |
| **RMSE** | 0.0208 | 0.0203 | 0.0207 |
| **MAPE** | 0.2370 | 0.2335 | 0.2392 |

Table D2: Score Comparison for XGBoost Distance (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6030 | 0.6403 | 0.6495 |
| **MSE** | 0.0439 | 0.0395 | 0.0411 |
| **RMSE** | 0.0182 | 0.0174 | 0.0177 |
| **MAPE** | 0.2096 | 0.1988 | 0.2028 |

Table D3: Score Comparison for XGBoost One Mile (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5352 | 0.5496 | 0.5532 |
| **MSE** | 0.0514 | 0.0495 | 0.0524 |
| **RMSE** | 0.2267 | 0.2224 | 0.2290 |
| **MAPE** | 0.0199 | 0.0194 | 0.0198 |

Table D4: Score Comparison for XGBoost Three Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.5885 | 0.5993 | 0.6086 |
| **MSE** | 0.0455 | 0.0440 | 0.0459 |
| **RMSE** | 0.2133 | 0.2098 | 0.2143 |
| **MAPE** | 0.0185 | 0.0183 | 0.0185 |

Table D5: Score Comparison for XGBoost Five Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.6314 | 0.6838 | 0.6957 |
| **MSE** | 0.0408 | 0.0347 | 0.0357 |
| **RMSE** | 0.2019 | 0.1864 | 0.1890 |
| **MAPE** | 0.0172 | 0.0160 | 0.0163 |

**Group E: Ridge Regression Score Comparison**

Table E1: Score Comparison for Ridge Baseline (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3494 | 0.3276 | 0.3613 |
| **MSE** | 0.0720 | 0.0739 | 0.0750 |
| **RMSE** | 0.2683 | 0.2718 | 0.2738 |
| **MAPE** | 0.0237 | 0.0239 | 0.0242 |

Table E2: Score Comparison for Ridge XGBoost Distance (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3776 | 0.3571 | 0.3883 |
| **MSE** | 0.0688 | 0.0706 | 0.0718 |
| **RMSE** | 0.2624 | 0.2657 | 0.2679 |
| **MAPE** | 0.0230 | 0.0233 | 0.0236 |

Table E3: Score Comparison for Ridge One Mile (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.3974 | 0.3693 | 0.3952 |
| **MSE** | 0.0667 | 0.0693 | 0.0710 |
| **RMSE** | 0.2582 | 0.2632 | 0.2664 |
| **MAPE** | 0.0228 | 0.0232 | 0.0236 |

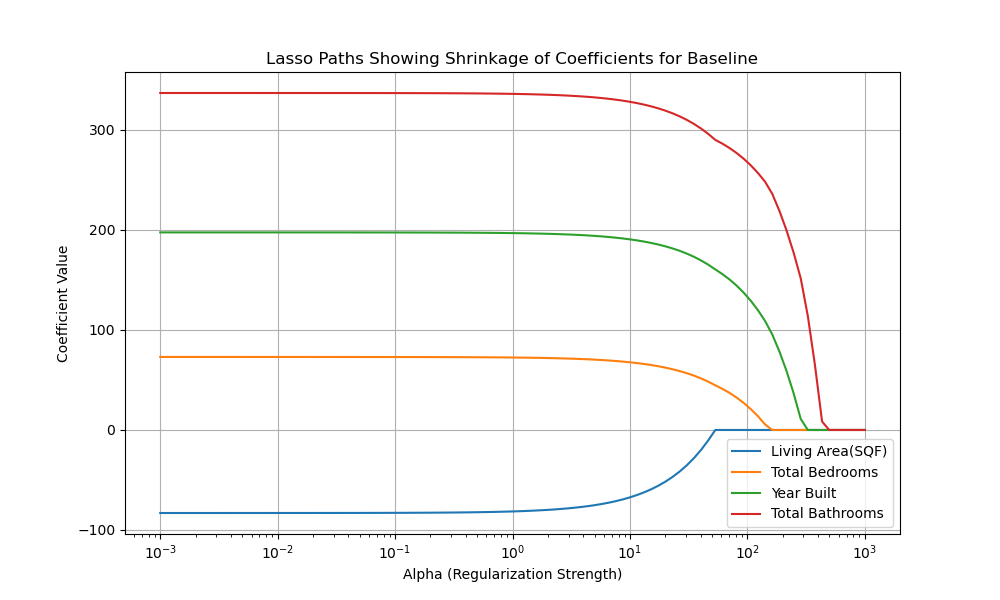
Table E4: Score Comparison for Ridge Three Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4563 | 0.4342 | 0.4674 |
| **MSE** | 0.0601 | 0.0621 | 0.0625 |
| **RMSE** | 0.2452 | 0.2493 | 0.2500 |
| **MAPE** | 0.0216 | 0.0221 | 0.0222 |

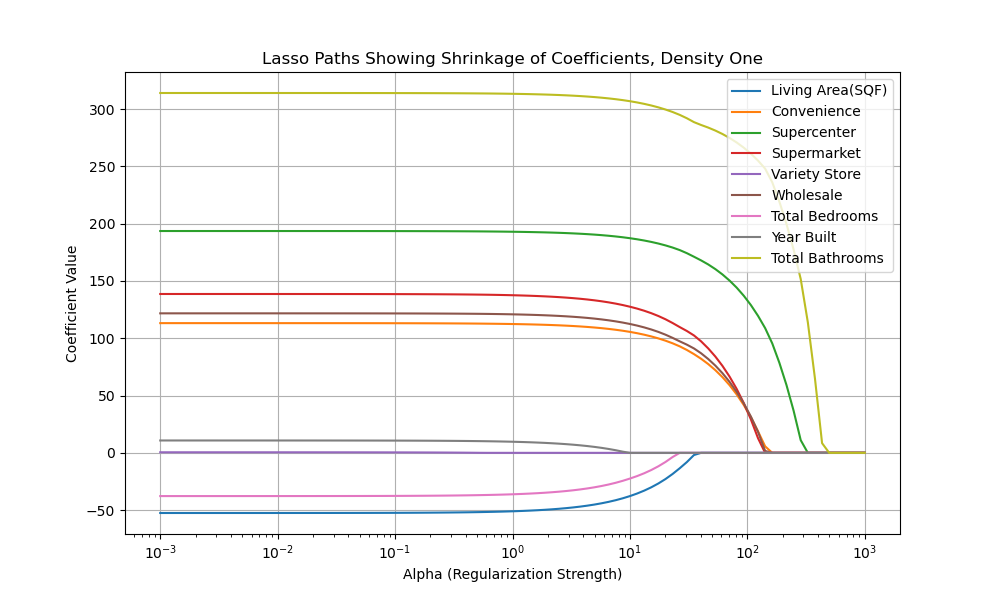
Table E5: Score Comparison for Ridge Five Miles (Log)

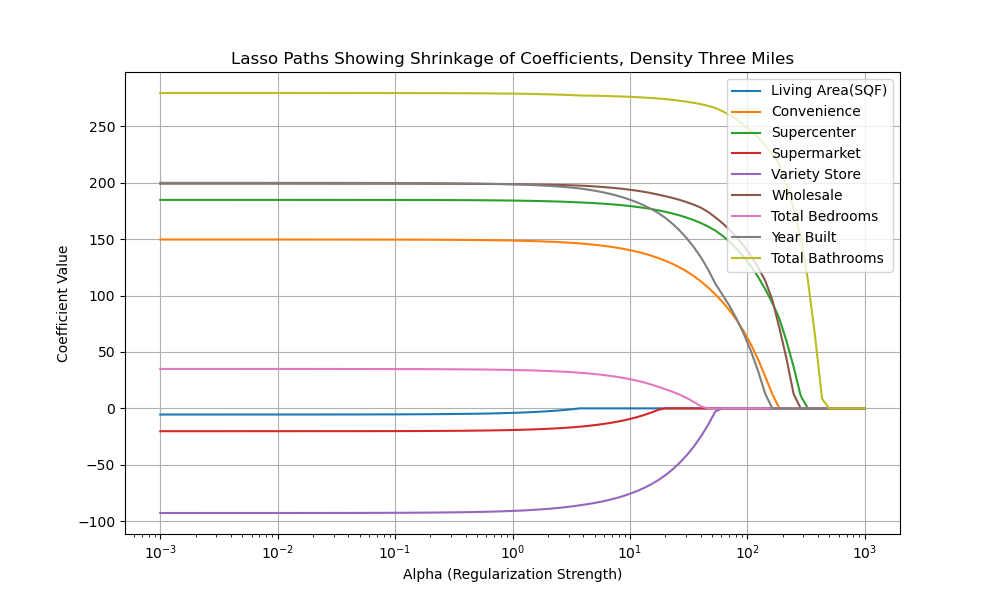
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4681 | 0.4572 | 0.4894 |
| **MSE** | 0.0588 | 0.0596 | 0.0599 |
| **RMSE** | 0.2426 | 0.2442 | 0.2448 |
| **MAPE** | 0.0211 | 0.0216 | 0.0219 |

**Group F: Lasso Regularization Graphs**

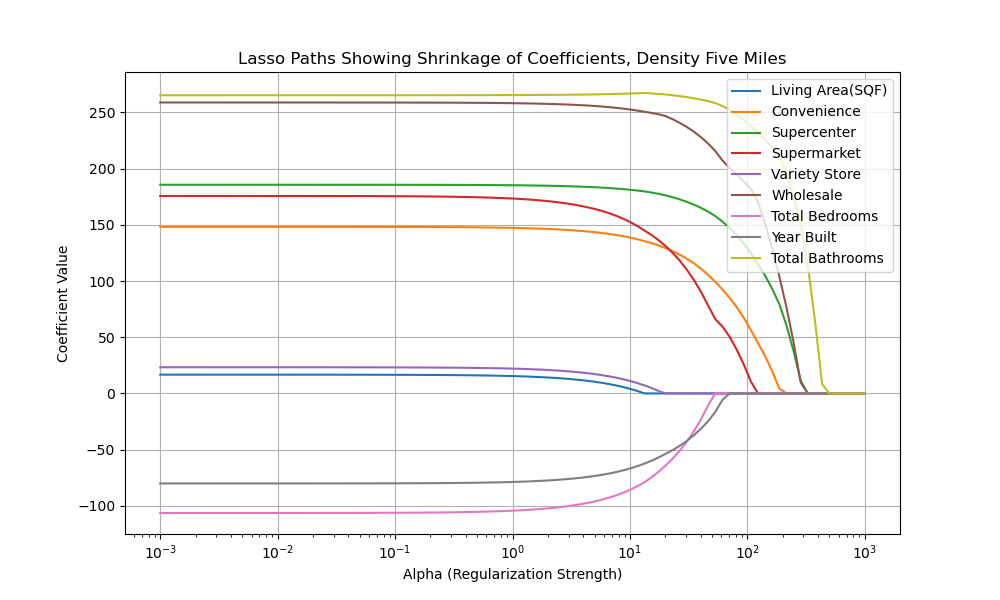
****

(Fig. F1: Lasso Regularization Graph for Baseline Features)

(Fig. F2: Lasso Regularization Graph for Density One Mile Features)



(Fig. F3: Lasso Regularization Graph for Density Three Features)



(Fig. F4: Lasso Regularization Graph for Density Five Features)

**Bibliography**

Antipov, Evgeny A., and Elena B. Pokryshevskaya. “Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a Cart-Based Approach for Model Diagnostics.” *Expert Systems with Applications* 39, no. 2 (February 2012): 1772–78. https://doi.org/10.1016/j.eswa.2011.08.077.

Bin, Okmyung, Tatiana Filatova, and Koen de Koning. “Improved Methods for Predicting Property Prices in Hazard Prone Dynamic Markets.” *Environmental and Resource Economics* 69, no. 2 (October 26, 2016): 247–63. <https://doi.org/10.1007/s10640-016-0076-5>

Brown, Joseph D. “Determinants of Loyalty to Grocery Store Type.” *Journal of Food Products Marketing* 10, no. 3 (September 21, 2004): 1–11. [https://doi.org/10.1300/j038v10n0301](https://doi.org/10.1300/j038v10n03_01).

Cerrato Caceres, Belkis, and Jacqueline Geoghegan. “Effects of New Grocery Store Development on Inner-City Neighborhood Residential Prices.” *Agricultural and Resource Economics Review* 46, no. 1 (April 2017): 87–102. <https://doi.org/10.1017/age.2016.29>.

Chamhuri, Norshamliza, and Peter J. Batt. “Factors Influencing the Consumer’s Choice of Retail Food Store.” *Stewart Postharvest Review* 5, no. 3 (2009): 1–7. <https://doi.org/10.2212/spr.2009.3.1>.

Emerson, Robert E., Lisa House, and Marco A. Palma. “Evaluating Factors Influencing Grocery Store Choice.” American Agricultural Economics Association Annual Meeting, Montreal, Canada, *American Agricultural Economics Association*, July 27-30, 2003, n.d. 10.22004/ag.econ.21909.

Esri. (2019). The Impact of Location on Property Value. Esri White Paper.

Flanagan, Albert, and Steven Peterson. “Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal.” *Journal of Real Estate Research* 31, no. 2 (January 1, 2009): 147–64. https://doi.org/10.1080/10835547.2009.12091245.

Kiel, Katherine A., and Katherine T. McClain. “House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation.” *Journal of Environmental Economics and Management* 28, no. 2 (March 1995): 241–55. <https://doi.org/10.1006/jeem.1995.1016>

Law, Stephen, Brooks Paige, and Chris Russell. “Take a Look Around: Using Street View and Satellite Images to Estimate House Prices.” *ACM Transactions on Intelligent Systems and Technology* 10, no. 5 (September 9, 2019): 1–19. https://doi.org/10.1145/3342240.

Leibtag, Ephraim S. “Where You Shop Matters: Store Formats Drive Variation in Retail Food Prices.” *Amber Waves: The Economics of Food, Farming, Natural Resources, and Rural America* 3, no. 5 (November 2005): 13–18. <https://doi.org/10.22004/ag.econ.127354>.

Orford, Scott. “Modelling Spatial Structures in Local Housing Market Dynamics: A Multilevel Perspective.” *Urban Studies* 37, no. 9 (August 2000): 1643–71. <https://doi.org/10.1080/00420980020080301>

Shahzad, Mohsin, Abaid Ullah Zafar, and Hassan Zulqarnain. “Factors That Affect The Choice Of Consumers In Selecting Retail Store, For Grocery Shopping.” *International Journal of Multidisciplinary and Current Research* 3, no. Nov/Dec 2015 (November 20, 2015): 1167–72. <https://doi.org/10.2139/ssrn.2584357>.

Stegelin, Forrest. “Is Being Big Better? Shoppers Compare Food Merchandisers.” *Journal of Food Distribution Research* 47, no. 1 (March 2016): 81–82. <https://doi.org/10.22004/ag.econ.232317>.

Tomal, Mateusz. “Modelling Housing Rents Using Spatial Autoregressive Geographically Weighted Regression: A Case Study in Cracow, Poland.” *ISPRS International Journal of Geo-Information* 9, no. 6 (May 26, 2020): 346. https://doi.org/10.3390/ijgi9060346.