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

|  |
| --- |
| ***Abstract****: This research paper aims to investigate the impact of Points of Interest (POIs) on residential property valuation using machine learning techniques. The question is whether certain metrics, such as proximity and density, to a POI externality can significantly influence property values. The factor chosen was grocery stores, a type of shopping center typically dependent on location. With a focus area being on the Atlanta, Georgia Metropolitan Region, this paper finds that distance and density metrics are not only significant, but also substantive. This paper further explores the types of grocery stores and their influences, to see which factors most significantly influence property values. The advantages and disadvantages of the different machine learning methods used are discussed, therefore providing insights into effective feature engineering strategies in the context of real estate valuation models.* |

**Introduction**

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

What determines a property’s value? Conventional wisdom often places the blame on market dynamics—the ever-changing conditions based on supply and demand, interest rates, and economic stability. Still, property valuation, or the assessment of the value of a property, is usually more nuanced, taking into many factors, each having varied levels of an impact. Internal factors like square footage, age, condition, and features like the number of bedrooms, bathrooms are also key components in determining a property’s price. For example, as a baseline, it can generally be said that, as square footage increases, so does sale price (Bin, Filatova, Koning 2016, 253-54).

However, what of external factors? The amenities surrounding locations that buyers have to assess in determining if the area is where they want to live. Parents might want schools to be closer so that their children could walk there, individuals without a personal vehicle might want a home closer to a transportation hub (busses, trains, subways, etc.), and others might be more interested in parks or employment hubs. It has been shown that these types of amenities affect housing prices (Esri, 2019; Orford, 2000). Other externalities, such as proximity to industrial areas or airports, may actually reduce property values (Kiel and McClain 1995, 247-53).

With the advent of big data and machine learning, there is an opportunity to more precisely quantify the impact of these POIs on residential valuation. Real estate appraisal can possibly be determined with machine learning algorithms. Normally, an ordinary least squares (OLS) regression is used, but the problem with an OLS model is that it is highly sensitive to outliers and can often suffer from severe multicollinearity (Tomal 2020, 346). Therefore, other methods, like gradient boosting or Random Forest, should be pursued that can predict property prices with higher accuracy than traditional hedonic pricing models (Antipov and Pokryshevskaya 2012; Flanagan and Peterson 2009).

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

The geospatial analyst would look at the qualities of the neighborhood and quantify it, measuring the environment as influencers as a way to go hand-and-hand with the internal features. For example, satellite imagery can be used to predict housing prices. In fact, an attempt was done by Stephen Law et al., where the neighborhood was shown to be a more predictive factor in housing prices than the local streets and the houses themselves (Law, Paige, Russell 2019, 16). But what POI-derived features should we be looking at as the most predictive of residential property values?

Whatever amenity it may be, there is always one that stands out as an essential service among the others: the convenience factor—grocery stores and hybrid shopping centers that carry essential goods and services that are required for modern day living. In other words, without modern grocery shopping centers, people wouldn’t have this convenience factor and that would severely hamper their every day functioning. At best, a home owner would have to shop at a local farm, at worst, they would have to travel for miles to get essential supplies. And that distance to the nearest grocery store can be a major factor in determining housing price.

Therefore, we want to see how grocery stores influence current listing price, just like any other external amenity. And because these are external amenities, they can both be categorized and quantified, allowing for machine learning models to be implemented. The next section provides an explanation as to why the categories for our models were chosen, as well as support for the so-called convenience factor. We will then present the basic research methodology, which additionally uses a custom-made Javascript HTML app that utilizes Overpass API. The design will answer:

1. Which POIs are most predictive in regards to current listing price?
2. Do non-traditional or traditional grocery stores matter more in predicting housing costs?
3. Do grocery stores actually have an impact in comparison to the baseline?

Once the data is analyzed using various machine learning methods, like the suggested Random Forest, we will discuss further optimization and the potential pros and cons of using these methods.

**Literature Review**

**Categorization**

It has already been demonstrated that other external amenities at least have “some” influence on housing price. We also know that this effect can be calculated via hedonic pricing models, machine learning models, and through geospatial analysis. So, how exactly would we classify a grocery store and quantify it? Ephraim Leibtag (2005) outlined a base framework for what the “grocery store” categories could be, he split them into the traditional grocery stores, which were the “conventional supermarkets”, specialty stores, warehouse stores, and the like, all limited in their size. Then, there are the non-traditional grocery stores, which were supercenters, wholesale clubs, and other mass merchandizers that were not only large, but were hybrid in their goods sold.

Forrest Stegelin (2016) had a similar classification, narrowing the search to the North Georgia and the Atlanta Metro counties. He defined the various stores into regional chains, warehouse clubs, national chains, mass merchants, and department stores. However, the examples given coincided with Leibtag’s (2005) classifications. For example, the warehouse clubs are the wholesale clubs, thanks to examples like Costco and Sam’s club, and the mass merchants are the supercenters, as they both include Walmart and Target.

Emerson et al. and Joseph Brown (Emerson, House, Palma 2003; Brown 2004) proposed a different type of categorization, splitting the grocery stores into three different types based on their quality, price, and services provided. Generally, there were the “higher price, higher quality” Type A stores, the medium quality, medium price Type B stores, and the low quality, low price Type C stores. The issue is that this classification does not mesh well with Leibtag’s and Stegelin (2016) categories. There isn’t even a “non-traditional” or “traditional” filter. The main argument is that this classification makes it easier to find correlations between customer patterns, instead of segmenting the stores themselves.

We chose a more combined approach, requiring a traditional or non-traditional classifier, but also dividing grocery stores with quality and services. Palma et al. (2003) compares the Type stores to Stegelin’s (2016) mass merchants, namely Walmart, and thus Leibtag’s (2005) supercenters. Therefore, our more eclectic approach is still attempting to keep with the framework that the literature laid out.

**Grocery Stores as a Determinant**

A home owner would typically choose to shop at a grocery store based on its “utility”, defined as its convenience (distance from the home) and then its “assortment”. Assortment is defined differently based on the literature. For example, Brown (2004) went by variety, quality, and price, while Mohsin Shahzad et al. (Shahzad, Zafar, and Zulqarnain 2015) argues in favor of “prices, quality of products, and branding”. While all factors determine store choice, it was found that location was a statistically significant influencer on consumer choice with more than half of respondents preferring to shop “within their residential area”. Respondents who preferred supercenters/wholesale stores were only moderately affected by location. (Shahzad, Zafar, Zulqarnain 2015, 1170- 71).

Interestingly, Brown argued that store loyalty might occasionally surpass distance as a factor, but then says that distance can cause an “inertia”, causing individuals, especially those with a low income, to default to closer stores Brown (2004, 2-3). Palma et al. (2003, 9) adds that, while income of the family can determine choice of grocery store Type, “education, gender, and age appear to have little systematic effect on store choice.” This hints of the further importance of grocery stores over other external amenities, and possibly diminishes some worry of certain other factors manipulating results.

In keeping with Brown’s (2004) conclusion that price affects store choice, Leibtag (2005) also claimed the same, further arguing that non-traditional stores had consistently lower prices than those of traditional ones. Therefore, it should be expected that lower income families would be more likely to choose non-traditional stores. Stegelin (2016, 81-82) gave mixed results to this kind of question. He said that supercenters, like Walmart, actually received subpar scores compared to other grocery chains in the southeast region, including those having comparable prices, with regional supermarkets ranking highest. He further claims that shoppers, regardless of income, had a tendency 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 also reinforce Brown’s (2004) point about income, saying that demographic characteristics matter, and once again says that higher income earners are able to change location. Lower income groups, instead, value the “traditional retail formats”. Ultimately, "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 that have access to cars. Alternatively the "traditional" retail format is chosen due to closer proximity, as lower income individuals have limited storage capability and can't afford travelling costs.

A study in the northeast region done on supercenters (Target and Walmart) attempts to see if these conclusions make sense. Supercenters would be classified as nontraditional shopping locations that are supposed to offer lower prices, but Stegelin (2016) said the prices didn’t really matter in comparison to other chains. According to the results, when the total distance was 800 meters, on average there was a decrease in sale price, with 400-599.9 experiencing the largest decrease. However, all other splits (0-199.9, 200-399.9, and 600-800) experienced an increase in sale price. It should be noted all supercenters were located in poor neighborhoods (Caceres and Geoghegan 2017).

**Conclusion: Proximity or Density?**

From the literature, it can generally be asserted that distance from grocery stores has some kind of an effect on sale price. Additionally, it would appear that the type of grocery store matters as well, as certain consumers either can’t afford to go to a certain type of store, prefer variety in their selection, prefer certain “brands”, or may be forced to choose a type of store based on its distance from their home.

Most of the research seemed to focus on proximity, usually classified as “proximity to the closest POI”, as the feature to be focused on. It would seem that low income individuals are forced to rely on distance as a factor, but the data is inconclusive on what type of store would actually be encouraging this. Interestingly, it would seem that development trends spark rental price increases. Considered “revitalization”, even in poor neighborhoods, the development of new grocery stores shows a positive increase in rental prices. This applies most to those houses closest to the new lots.

Therefore, we propose another feature in addition to proximity: density. We also know that retailers format their locations in central locations. Therefore, it should be expected that these central locations, or locations where there are numerous amounts of grocery stores, should influence housing prices in a positive way, perhaps even more so than distance. This is because density more tells of urbanization and development, where if one went by distance in a highly urbanized region, that same distance is probably duplicated multiple times because of the sheer number of stores in a radius around the home. We would hope, then, to compare density against distance to see which would produce a stronger model, as well as which, overall, produces greater predictive factors.

**Methodology**

**Research Design**

Using a combination of influences from the literature, we decided to base our design around the main target, which was appraisal prices of houses. The southeast region, specifically Atlanta, Georgia, seemed to be a “hotspot” for some of the literature, so the Atlanta Metropolitan Statistical Area was chosen to be used for the housing data. The housing data itself contains features that can be expected of a typical region report, namely characteristics like sale price, size, age, number of bedrooms, and their geolocation coordinates.

Continuing with the combined approach, we decided to create a mixture of categories for the grocery store “features”, dividing them into non-traditional and traditional types (Stegelin 2016, 82). The traditional features are:

1. Supermarkets: A full service grocery store that often sells a variety of non-food products as well. Almost always part of a chain (Publix, Harris Teeter, Piggly Wiggly, BiLo, Ingles, Bells, Earthfare).
2. Variety Store: A variety store or price-point retailer is a retail shop that sells inexpensive items, usually with a single price point for all items in the store (Dollar Tree, Family Dollar, Dollar General).

And then the non-traditional:

1. Warehouse Club: A non-traditional membership-retailer hybrid that sells bulk products in a warehouse environment. At least 40% of products devoted to grocery (Costco, Sam’s Club, BJ’s).
2. Supercenter: A non-traditional food and drug store combination that sells mass merchandise. At least 40% of products devoted to grocery. At least 40,000 square feet. (Target, Walmart).
3. Convenience Store: A non-traditional, limited store that sells variety of general merchandise, including packaged food products (7-Eleven, Quick Trip).

Before we could even make the models, we had to collect the data for the distance-based features, as well as the density features. We followed by example and chose to define distance as “nearest POI”, but density was a bit more difficult because of Caceres and Geoghegan’s (2017) inconclusive results. Therefore, we chose to split density between a one mile radius around the desired location, a three mile radius, and a five mile radius. With the premise completed, we could then assert the hypothesis we’re testing against, which is that neither distance nor density has any effect on housing price.

We were immediately presented with an issue of “collecting” the data for the proposed features. Normally, one would use a program like OpenStreetMap or Google Places API, but the requirement was to ensure all property and POI data were geocoded for spatial analysis, as well as perform spatial joins to link properties with nearby POIs. This is not done for us using a standard API. So instead, we created an app through Javascript that utilized Overpass, in order to extract both the distance values and the density values based on POI type, or in certain cases “brand name”.

Once all data was collected, missing values, outliers, and inconsistencies in the datasets had to be dealt with, and finally the main design could be proposed. Since Antipov and Pokryshevskaya (2012) demonstrated that a Random Forest model can be used for this type of design, we chose that as our initial model. A Random Forest model combines multiple decision trees, averaging the predictions on subsets of data to make predictions. It is able to capture non-linear relationships and interactions between amenities and rental prices. The problem is typically that Random Forests can struggle with imbalanced data, biasing towards certain features and yet still having a high computational cost.

Therefore, we added on an XGBoosting model as well. An XGBoost model is a type of gradient-boosting model that builds the decision trees sequentially, each iteration improving on the last as it accounts for errors. XGBoosts are faster, more powerful, and better at handling larger datasets, including those with missing values, which housing data tends to have. Since we didn’t just want to see feature importance, but also correlations as well, we wanted to use regression and regularization techniques in addition to the tree methods.

Multiple linear regression is the most basic regression that can be used, as it cannot be tuned, so Lasso and Ridge regression models were also used. A Lasso model shrinks the coefficients to nearly zero, while the Ridge regression shrinks the coefficients towards zero, but never to exactly zero. These methods allow us to understand the linear relationship between rental price and external amenities along with other features.

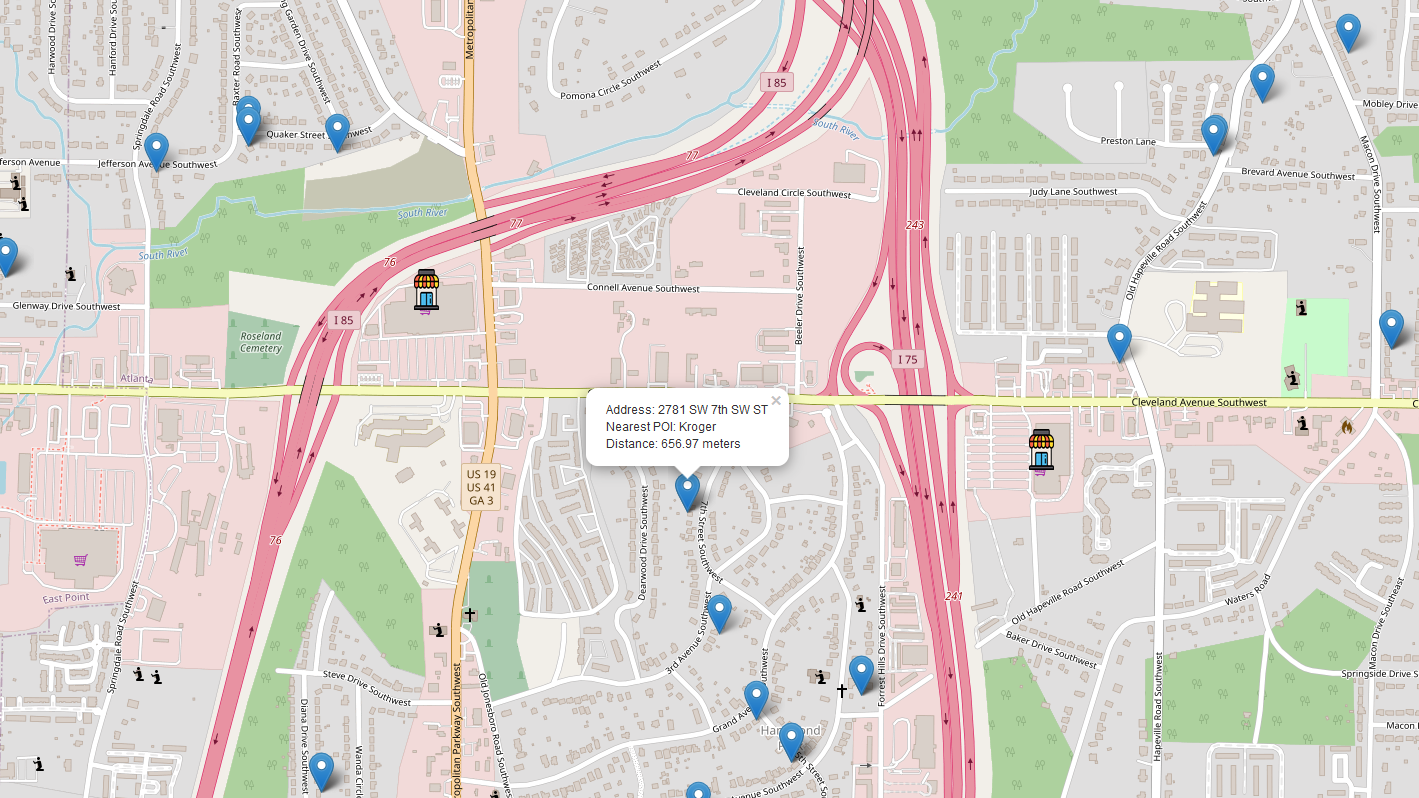
Once the models were decided, we used the XGBoost and Random Forest models to find feature importance as well as partial dependency, which shows the relationship between the selected feature and the target, while averaging out the effects of all other features. The regularization techniques allowed us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

**Using Leaflet: Data Collection**

To gain the data for the density and distance values, an app was created using Javascript as a base. It relies on Leaflet CSS, an open-source library for interactive maps. The map in question being used was OpenStreetMap API, a geographic database available to anyone, and can be updated by the community at large. Since geospatial data is being used, the original housing data had to be converted into a geogson format so it can be read be Leaflet. Within the Javascript, the boundaries were then set for the Atlanta area.

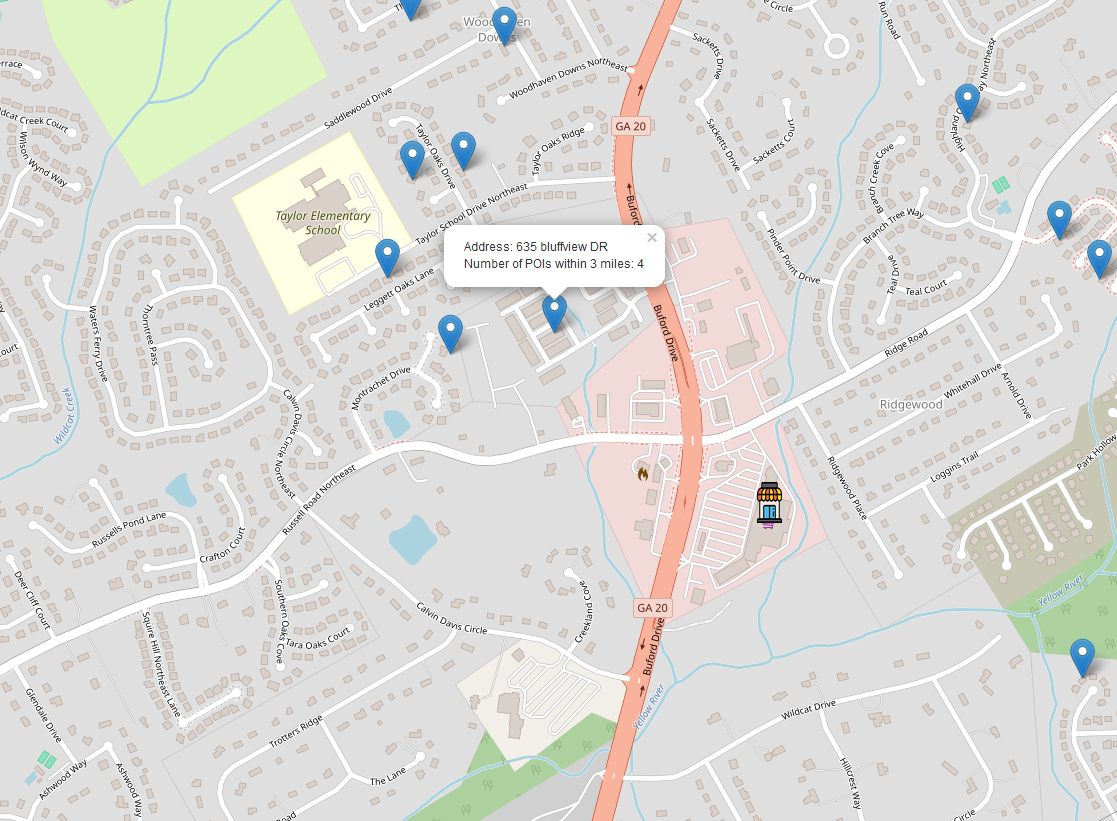
Calculating distances was fairly straight forward, done in one function. Essentially, assuming the POI had an address, the closest POI was reported, which already has its own function in HTML. This was all then recorded back into a CSV file for later use. Overpass uses a tag-based system, where the POIs, such as gas stations or parks, are classified under categories based by the features that they have. However, it can also go by brand-name if the store doesn’t fit a particular category. We based our traditional and non-traditional features so they would work well with Overpass’s tag-based system.

Fig. 1 shows the distance-based map for a particular address near Cleveland Ave SW, Atlanta, GA. When one clicks on a blue location marker, it will display the address, the nearest POI’s name (in this case we were searching supermarkets, so it found a Kroger to be the nearest POI), and then the distance to that nearest POI. The tags we searched were “shop=supermarket”, “shop=variety\_store”, “shop=wholesale”, “shop=convenience”, and then the brand names for what could classify as a supercenter, which were Target, Walmart, and technically Amazon and Aldi, but there weren’t enough of them in the region to be relevant.



(Fig. 1 of distance-based Overpass search of what would be downloaded to a CSV and used as data. One may 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 there is a Walmart, which is classified as a supercenter. One can exclude certain brands from Overpass searches, or exclude it in the code entirely.)

The density-based code had a similar function, except instead of focusing on the closest point, the function was based on the number of POIs within a radius that was defined in meters (based on a simple conversion from one, three, or five miles). After a radius was defined, the app mostly works the same. Fig. 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 point that out, and then display the distance to that. However, since this is the density-based model, which we selected the radius as three miles, it is showing the number of supermarkets within three miles, most of which are outside of this level of zoom.



(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 done for distance and the radius for one mile, three miles, and five miles, the data can then be combined with the original set based on what was being searched. Specifically, there would now be twenty features based on the five we outlined, each grouped by whether it’s distance-based or density-based. And if the data is density-based, it has to be split between the three different radiuses.

**Data Analysis: Scores**

Which type of grocery store is most predictive of residential property values, if at all? There are two rationale in establishing the alternative hypothesis. The proximity to certain amenities can influence property desirability and, based on density, different radii can capture local and broader neighborhood effects. A cross-tabulation was used for all models to test what 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 the model explains a large portion of the target's variability. We would want that score to be higher than the baseline, which is a model without any added features.

Mean squared error and root mean squared error are loss metrics that calculate the averaged squared difference between the predicted values and the target values. RMSE just squares the result, making it more interpretable. A smaller value would indicate better model performance and, unlike R2, these scores do evaluate predictive accuracy. Mean absolute percentage error measures the accuracy of a model's predictions relative to the actual values, with the error expressed as a percentage. For example, a MAPE is independent of a scale, with a smaller MAPE meaning better predictive accuracy.

Once the defaults were found, we tuned each of the models based on various parameters designed specifically around each model. Hyperparameter tuning allows for certain aspects of the cross-tabulation to be refined in an attempt to improve the score, such as the depth of the trees and the alpha value. We therefore used the “best parameters” to find the various desired results.

A train-test-validation split was used for the modeling, in order to account for overfitting and analyze all data. The training set is used for initial fitting, but isn’t necessary for a final metric. The validation set (20% of the data) is used for tuning the model and can help measure model performance. The combined train-validation set allows us to train on the full 80% of the data, therefore capturing more patterns. The test set is the main goal, any scores derived tells you 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 of an overfitting there is.

**Distance-Based Scores**

Only one set of results could be reported for the distance-based models, with each model having varying results. Table 1 displays the XGBoost scores for the tuned distance model. What it means is that there is slight overfitting, due to the 17.01% difference in the train-validation set and the test set. This was done intentionally, as we could have made the train-validation scores almost equal to the test scores, at the cost of the test scores being much lower. Therefore, for every model, it should be expected that the train-validation scores are going to be higher than the test scores.

Normally it is desired for the R2 to be higher, but it was stated that it doesn’t measure predictive power. The MAPE of 14.93% is telling us that the model's predictions deviate from the actual values by 14.93%. This is more desirable. The RMSE doesn’t tell us much by itself because the number is that high based on the target data we’re dealing with, and therefore we have to use it to compare with other models.

Table 1: 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 |

When it comes to other models, it was shown that the Random Forest iteration of the distance-based data was inferior in every way compared to the XGBoost model (see Appendix). The R2 was lower, and the MSE, RMSE, and MAPE were higher than their XGBoost counterparts. This will be the pattern for all upcoming density-based variations as well, which may indicate that the Random Forest model was not the best approach for this dataset.

Table 2 shows the multiple linear regression scores for the distance-based data. The MLR is an alternative approach to the XGBoost and Random Forest, and therefore it has different types of parameters to tune. Linear models are for smaller datasets and also assumes linearity. The linear models had the target log-transformed in order to improve performance. While it may appear that the scores are superior to the XGBoost, they are not. If the XGBoost model was also log-transformed, it would have better scores in every metric, if going by the test set against the Lasso.

Table 2: 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 had improved scores compared to the linear regression model. However, their scores were extremely similar to each other. This could be explained by how linear models are not designed for non-linear results nor can they handle the complexity that XGBoost models can. Table 3 shows the Lasso distance scores, which, visibly, only had its R2 be superior to the Ridge model (see Appendix). Once again, even though the Lasso and Ridge models did improve the scores, it didn’t do so enough to counter the distance (excluding the MAPE).

Table 3: 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 split into three different sets based on the radius. The one mile radius had the lowest number of POIs, simply by default. It was also the radius that had the most amount of locations that had zero POIs within the selected radius, which can severely affect the results. Table 4 demonstrates just how low the scores could be, lower than the distance model. Interestingly, despite having an inferior score compared to the XGBoost distance score, its log scores, including the MAPE, are actually superior to all distance-based linear models. As before, the Random Forest model is inferior in all ways to the XGBoost Model, although the scores are not low enough to be worse than the distance-based linear models (see Appendix).

Table 4: 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 5 once again shows that, of all the linear models, the Lasso is superior. Here, however, it’s a bit more apparent (see Appendix). The Lasso model has a higher R2 than the Ridge and a lower RMSE and MSE. Unlike the XGBoost, the one mile linear model is superior in all ways to the distance-based model. This doesn’t only apply to the Lasso, as each linear model for the density has higher scores in all categories compared to the distance models.

Table 5: 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 null hypothesis was that grocery stores do not affect housing prices. Because our alternative hypothesis, H1, is that housing prices are affected by their proximity to grocery stores, it also suggests that the number of grocery stores will affect housing prices. Therefore, it can be inferred that, as the number of grocery stores increases in proximity, so would the housing price. So, if we expand the radius, it should be expected that the stores will become more relevant as features.

The three mile radius demonstrates just how much of an effect increasing the radius can have. Table 6 and 7 shows the overall increase in all scores compared to the one mile scores, as we increase the radius. As always, the Random Forest had inferior scores, although they still improved in comparison to the one mile. Although only marginally, it also manages to surpass the distance-based XGBoost scores. The three miles density-based Random Forest model, in comparison to the Random Forest distance-based model, also had much higher improvement (see Appendix). Only its R2 was superior for the Lasso model, in comparison to the Ridge (see Appendix).

Table 6: 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 7: 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 8 shows once again a steady increase in scores, although not as drastic as the increase from one mile to three miles. The scores now clearly surpass the XGBoost distance scores. For the second time, the validation set has a higher score than the train-validation set, even though this model doesn’t have the highest percentage for overfitting. Table 9 also demonstrates the stalling in the large improvement, as it barely improves over the three mile model.

Table 8: 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 9: 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 |

The baseline is our confirmation of our null hypothesis. It is only the features as they were before the distance and density features were added. If one of our models fails to improve over the baseline, then that’s showing that the models failed to improve its predictive power and the new features are not useful. If the model is worse than the baseline, then the resulting complexity has added more noise without adding any predictive power. Luckily, according to Tables 10 and 11, it would appear that the baseline has overall lower scores than all of our models. While the R2 of the Ridge is higher than the Lasso, the Lasso was chosen instead for consistency’s sake (see Appendix).

Table 10: 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 11: 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 |

**Score Evaluation: Baseline**

Overall, the distance-based model managed to have higher scores than the density-based models until the three-mile marker. Interestingly, at no point did the distance-based linear models supersede their density-based counterparts. As expected, the density-based scores only increased as the radius increased in value, allowing for more POIs to be included. The XGBoost five-mile model was the best model of the twenty-five models. It had the highest score in every metric except for MAPE, assuming there was a log-transform. However, to make up for that, the five mile’s linear model’s MAPE was always superior to any other MAPE.

Because the distance-based models surpassed the one mile models, we can attempt to draw a few conclusions. One is simply that distance matters more, initially, when there are a fewer amount of grocery stores taken into account. It should be repeated that many of the houses had no grocery stores within one mile (of one category), and therefore distance would surpass its effect by default. Another possibility is that the two are more or less equal, and the comparison is perhaps unfair. It should be expected that most grocery stores, of one category, are within one or two miles of a home. Therefore, both the distance-based data and the one mile data, if it just has one store within one mile, are reporting the same location, just in an unorthodox way.

Either way, the baseline proved to have the worst overall scores in every category compared to all other models. We now know that predictive power increases the more POIs there are and we know that proximity matters. Is it safe, then, to reject the null hypothesis from these results? The answer is likely no, as all this told us is that there is a difference between models, but not necessarily if this difference is significant or meaningful.

Table 12 displays the results of a paired samples t-test. Specifically, it is the baseline model and the distance model and tests whether the two are statistically different. Judging from the results, every metric scored below the .05 threshold, therefore indicating a statistically significant difference. Further, Cohen’s d, the substantive significance, indicated 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** | 11.25157 | 0.0 | 3.55806 |
| **MSE** | 10.59367 | 0.0 | 3.35001 |
| **MAPE** | 15.80299 | 0.0 | 4.99734 |
| **RMSE** | 11.04426 | 0.0 | 3.49250 |

The same can be said for the multiple linear regression for the distance, perhaps to an even greater effect, as seen in Table 13. These patterns are then repeated when compared with both the density-one (Table 14) and the density-five models (Table 15). The density-one had the least improvement compared to the baseline, which is in-line with how the scores went.

Table 13: Significance Between Distance and the Baseline Ridge Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 24.85005 | 0.0 | 7.85828 |
| **MSE** | 18.41908 | 0.0 | 5.82462 |
| **MAPE** | 14.45024 | 0.0 | 4.56957 |
| **RMSE** | 19.83557 | 0.0 | 6.27256 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 6.19900 | 0.00016 | 1.96030 |
| **MSE** | 6.32123 | 0.00014 | 1.99895 |
| **MAPE** | 7.64451 | 0.00003 | 2.41741 |
| **RMSE** | 6.23715 | 0.00015 | 1.97236 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | 17.66182 | 0.0 | 5.58516 |
| **MSE** | 16.63609 | 0.0 | 5.26079 |
| **MAPE** | 45.22918 | 0.0 | 14.30272 |
| **RMSE** | 19.23333 | 0.0 | 6.08211 |

**Score Evaluation: Model Comparison**

However, we essentially already knew that the baseline was the worst model, and that distance and proximity are relevant. Therefore, the models should instead be tested against each other. Specifically, we want to know if the distance model is significantly different from the density models, and then we want to know if the density models are significantly different from each other. Table 16 shows us possibly what we’d expect. The mean of the first group, the distance, is greater than the mean of the second group for all metrics, the density-one.

Table 16: Significance Between Distance and Density One Mile XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-value** | **Cohen's d** |
| **R²** | 7.08791 | 0.00006 | 2.24139 |
| **MSE** | 6.19776 | 0.00016 | 1.95990 |
| **MAPE** | 9.23380 | 0.00001 | 2.91999 |
| **RMSE** | 6.83837 | 0.00008 | 2.16248 |

We knew from the score comparison that distance surpassed the one mile density. However, the density-three model marginally surpassed the distance model in the score comparison. Table 17 shows a reverse in course. Table 1 and Table 6’s scores were close, but Table 6 was superior. Table 17 shows this difference was actually much more significant than we originally thought. The t-statistic switches to the opposite direction in favor of density-three, and the results are statistically significant. However, as possibly expected, the effect size is very small. According to Table 18, the difference increases as the density increases, with the scores even surpassing that between the distance and density-one scores, just in the opposite way.

Table 17: Significance Between Distance and Density Three Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -3.58668 | 0.00587 | -1.13421 |
| **MSE** | -3.58233 | 0.00591 | -1.13283 |
| **MAPE** | -4.31045 | 0.00196 | -1.36308 |
| **RMSE** | -3.67372 | 0.00513 | -1.16173 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -7.32018 | 0.00004 | -2.31484 |
| **MSE** | -7.73862 | 0.00003 | -2.44717 |
| **MAPE** | -11.80155 | 0.00000 | -3.73198 |
| **RMSE** | -8.28376 | 0.00002 | -2.61955 |

We would also like to see if there is a statistically significant increase between each successive density model. Tables 19 and 20 show that, as the radius increases, the significance actually decreases (as in it gets closer to the threshold of .05). The effect size also gets weaker, and the difference between the values in the models do, as we predicted, begin to stagnate. This is all despite density-five having the best scores of all models.

Table 19: Significance Between Density One and Density Three Miles XGBoost Scores

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -10.19488 | 0.00000 | -3.22390 |
| **MSE** | -8.72040 | 0.00001 | -2.75763 |
| **MAPE** | -11.10359 | 0.00000 | -3.51126 |
| **RMSE** | -10.27136 | 0.00000 | -3.24809 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **T-Stat** | **P-Value** | **Cohen's d** |
| **R²** | -7.07683 | 0.00006 | -2.23789 |
| **MSE** | -7.75174 | 0.00003 | -2.45132 |
| **MAPE** | -10.94271 | 0.00000 | -3.46039 |
| **RMSE** | -7.76063 | 0.00003 | -2.45413 |

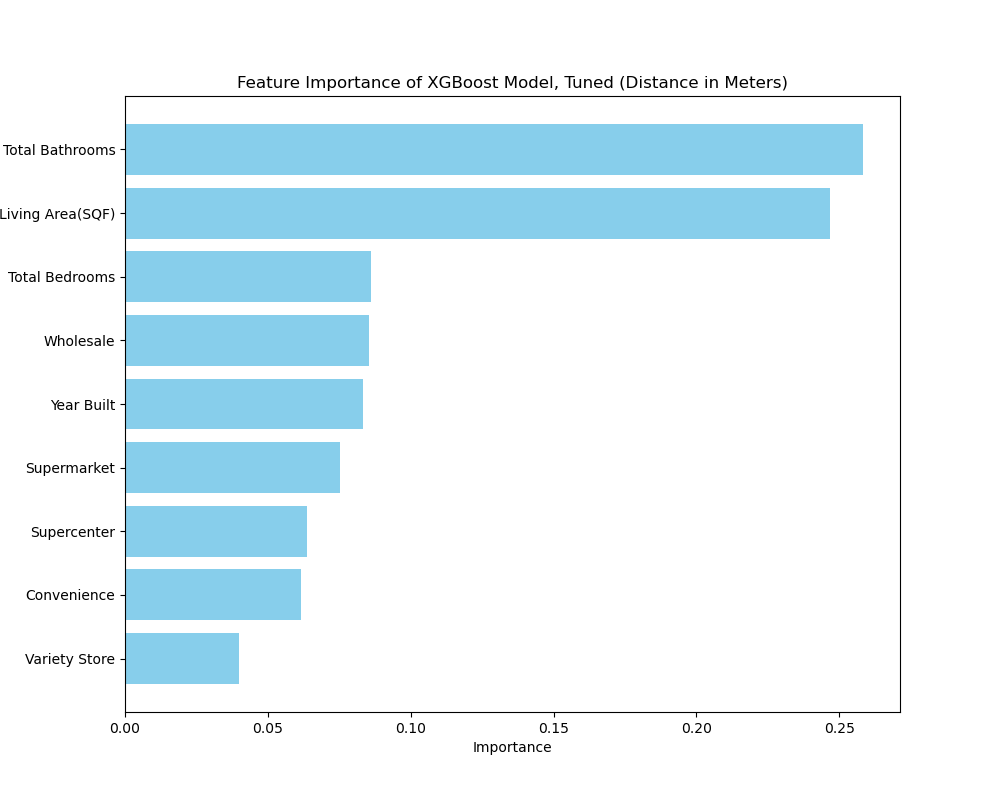
**Data Analysis: Feature Impact**

**Feature Importance**

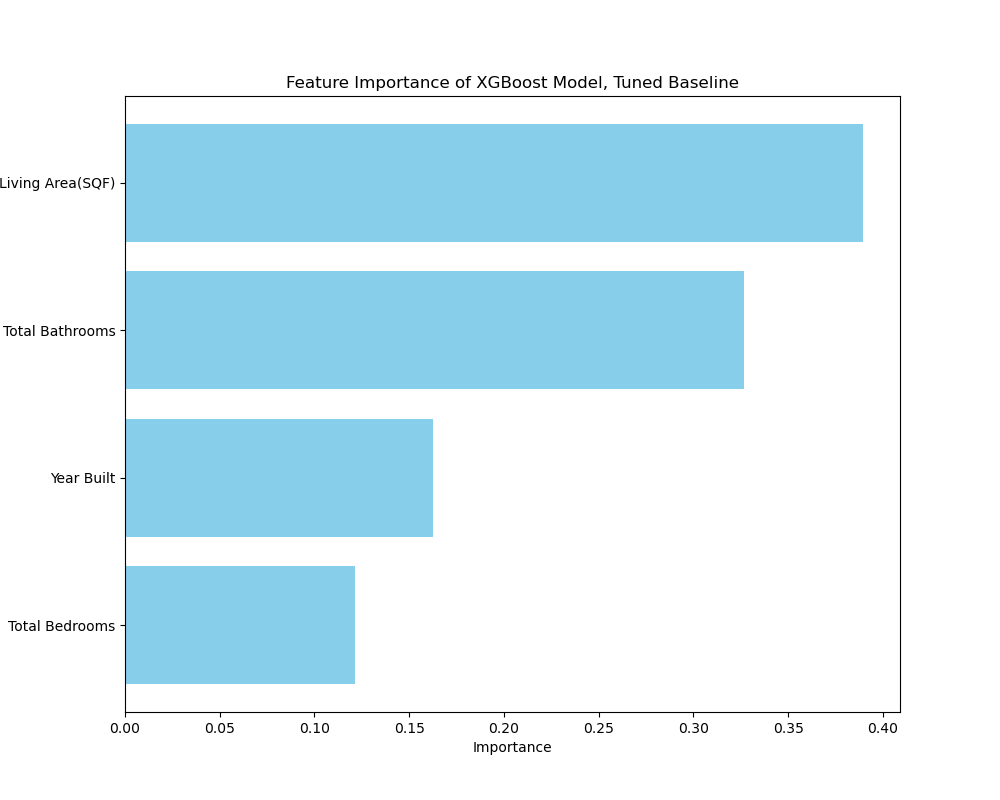
We now know if the models are significant, which models are superior to the others, and if they improve over the baseline. However, one of the goals was to see which features impacted housing prices. Feature importance is how much each feature contributes to the model's predictions. The Random Forest model and the XGBoost model computes feature importance differently, with XGBoost focusing on improvement over folds more, while the Random Forest focuses more on reduction in impurity over fold splits. Regardless, both can be used for feature ranking, and a high importance means that feature significantly affects performance.

Fig. 3 shows the feature importance for the XGBoost distance model. We can see that total bathrooms, living area, and total bedrooms are at the top. All of these are baseline features that manage to outperform the distance features in regards to feature importance. Wholesale is the first distance feature to appear, but then the final baseline feature is presented, followed by the rest of the distance features. Variety stores appear to have low feature importance, meaning that they contribute the least to the predictive power of the model.

Fig. 4 shows the feature importance of the baseline model, which relatively matches the layout of the distance feature importance. The main differences are that year built now surpasses bedrooms and the overall importance has increased. This isn’t surprising as the two had close values originally, and now multicollinearity is much less of a factor. It should be noted that in Fig. 4 that living area is now seen as the greatest predictor instead of total bathrooms. Only in the density models (Figs. 5 and 6) is living area of greater importance. However, in the Random Forest versions (see Appendix), living area always has the greatest importance.

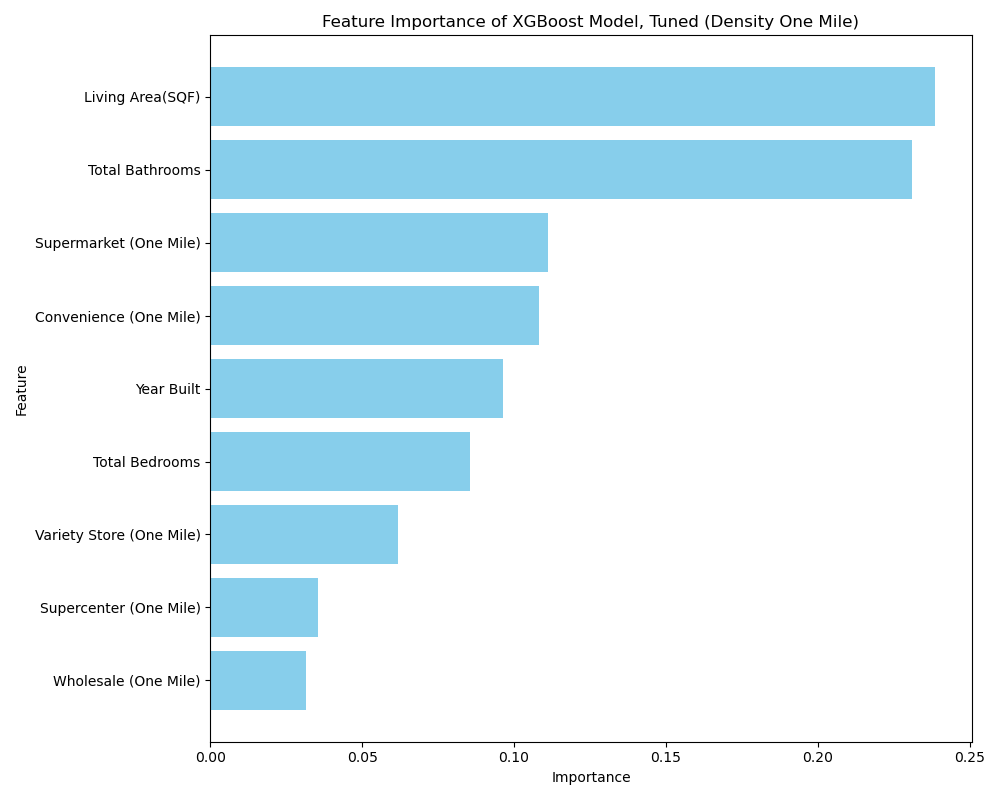


(Fig. 3: Distance XGBoost Model Feature Importance)

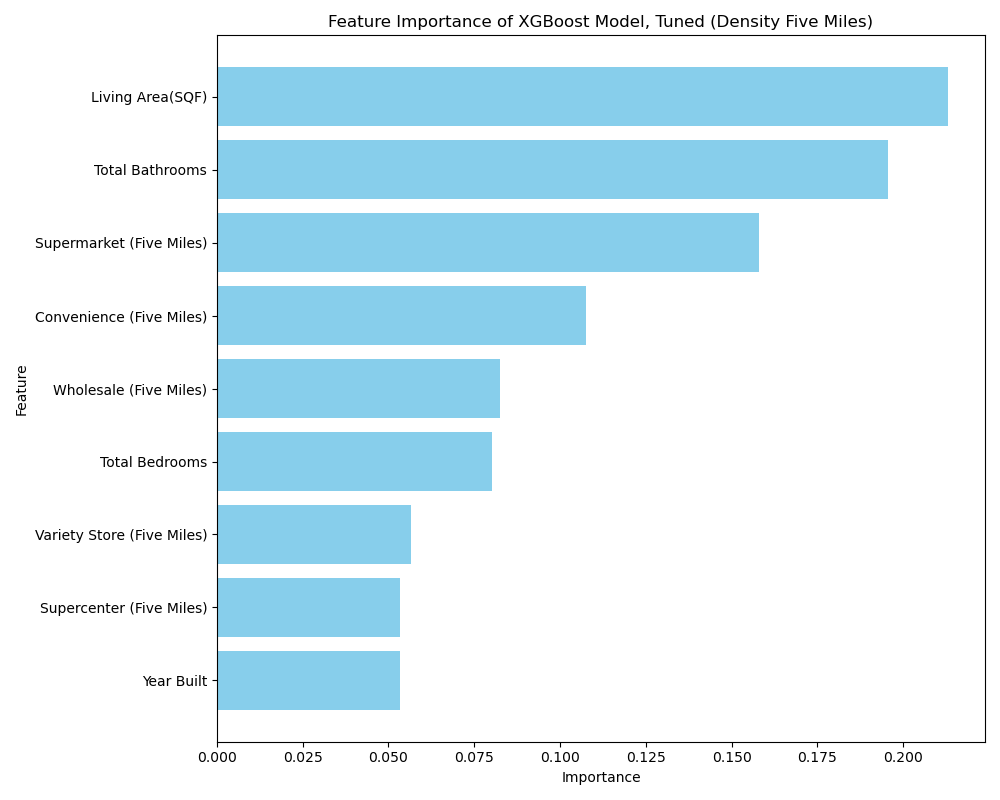


(Fig. 4: Baseline Model Feature Importance)

The density models give us clearer insights into the makeup of the data itself. Fig. 5 tells us that supercenters and wholesale stores not only have the least amount of predictive power, but according to the Random Forest version (see Appendix), wholesale doesn’t have influence at one mile radius. This is simply explained that there are not enough wholesale stores to make an impact in regards to predictive power. More importantly, we now see that supermarkets and convenience stores have actually managed to surpass baseline features even at one mile. In Fig. 6, this effect becomes even more profound, with the supermarket feature rapidly increasing its importance as the radius increases. It should also be noted that wholesale manages to surpass four other features at five miles, and in the Random Forest version supermarkets are considered the second most important feature (see Appendix).



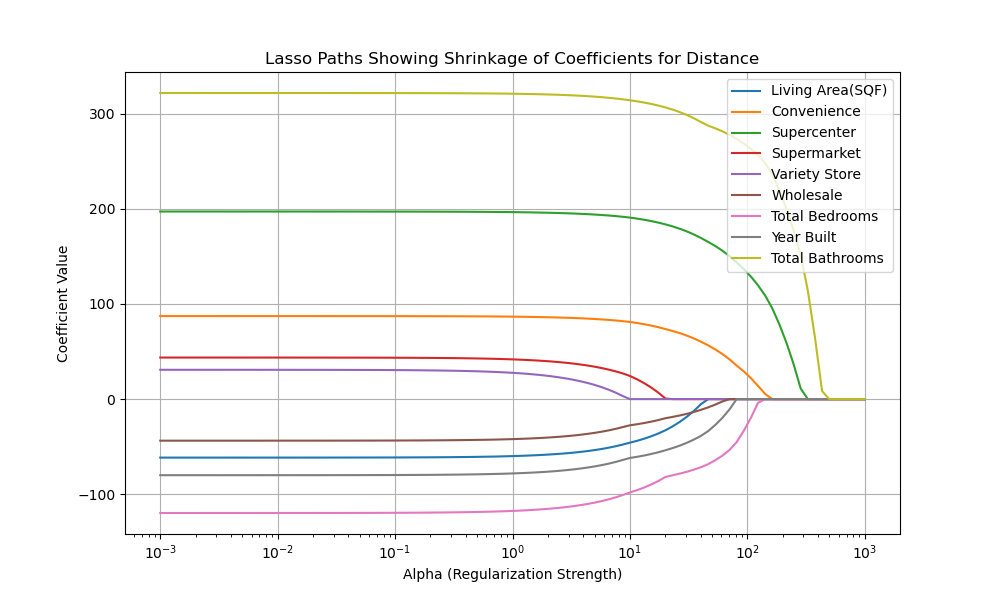
(Fig. 5: XGBoost Density One Mile Feature Importance)



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

Although normally the multiple linear regression would not have a feature importance display, one can show which features have a stronger influence when presented against a penalty on the size of the coefficients. Features with weaker or redundant relationships to the target variable are penalized more quickly, so their coefficients shrink to zero at lower alpha values. Features that manage to withstand the penalty longer typically have a stronger influence. It should be kept in mind that the linear regression models had much lower scores than models like the XGBoost and the Random Forest, so a difference in importance is expected.

This difference is evidence by just the distance model, as shown in Fig. 7. There is a complete reversal in course in comparison to Fig. 4. For example, although both begin with total bathrooms as the most important feature, the Lasso model goes completely in a different direction as to where it places features for influence. The Lasso regression is applying a penalty to the coefficients and assumes no regularization. Lasso graphs only consider direct linear effects, while the XGBoost reflects both linear and non-linear effects, thereby making some features appear more important than in Lasso. Furthermore, XGBoost accounts for interactions and is not dependent on the scaling that Lasso is. While the Appendix shows the density versions, it is unnecessary to compare the two as they are so vastly different from each other.



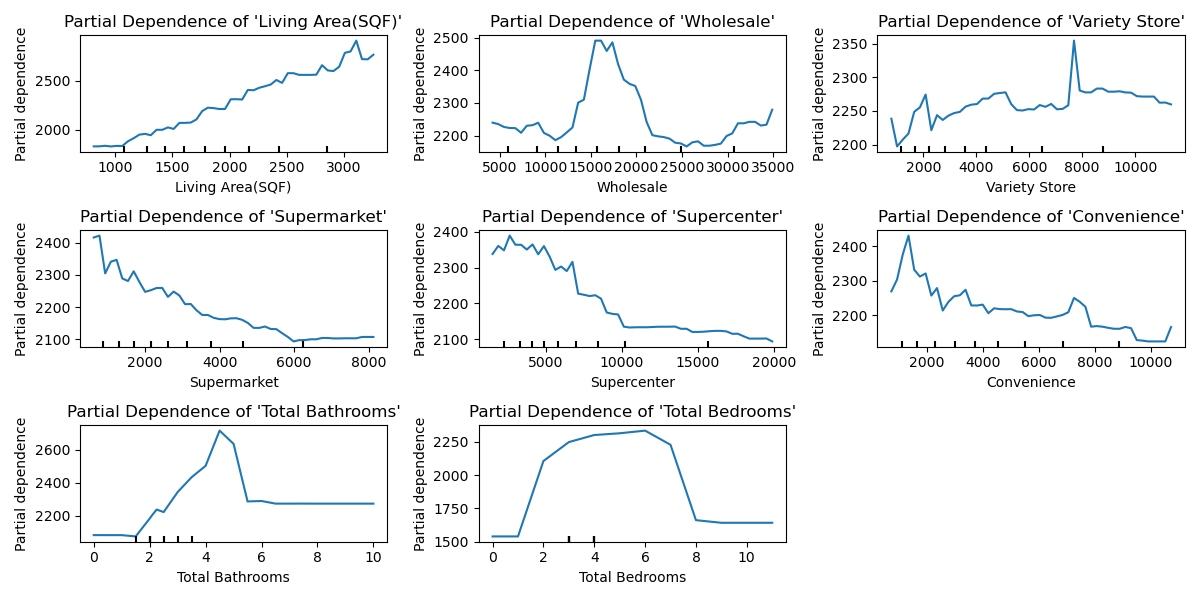
(Fig. 7: Lasso Regularization Graph for Distance Features)

**Partial Dependency**

While feature importance helps us get a vague assessment of predictive power, it gives us little insight on each feature’s impact on the actual target. We want to know predicted outcomes as the values change, which would give us an actual explanation as to why the wholesale feature was so low. A partial dependence plot (PDP) is a way of isolating features to see how certain features impacts a model's output. If a value increases/decreases, does the predicted outcome also increase/decrease in turn? Not only will we know how much a feature contributes to predictions, but the PDP will show us how it contributes, such as whether it is linear, exponential, or if it varies at certain thresholds.

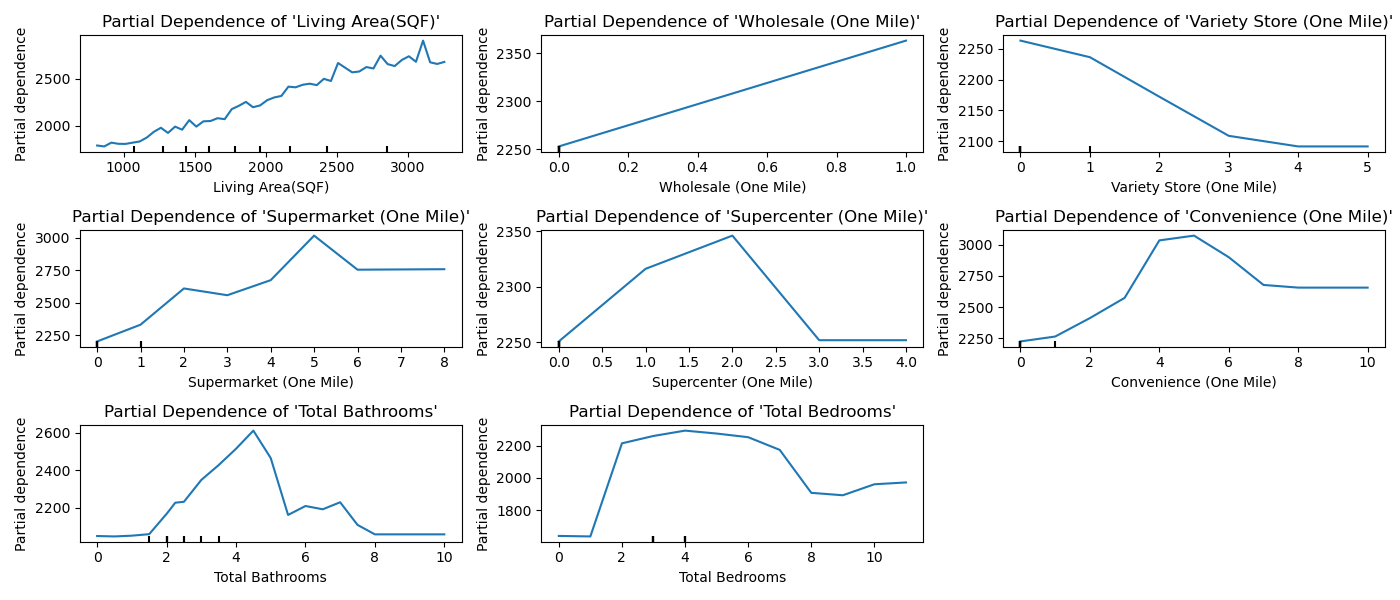
PDPs show features independently of each other, so a separate baseline “plot” is unnecessary. Therefore, Fig. 8 shows the PDP of the baseline features along with the distance features, when using the XGBoost model and the train-validation set. For these displays, the Y-axis represents the predicted outcome as influenced by the features plotted on the X-axis, in this case averaged housing prices. This allows us to see the change in predictions as the X values change. For example, as stated in the introduction, as living area increases, there is a rising trend in predictive values.

More importantly, we can see the opposite for three of the distance features, namely supermarkets, supercenters, and convenience stores. This means, that at least initially, housing prices have a higher predicted outcome the closer they are to these specific types of stores. The further away the closest POI becomes, the lower the predicted outcome. This is not the case for wholesale stores, which seems to hit an apex at around 9 miles (15,000 meters), before decreasing. This suggests wholesale stores in general have a tendency to be further away. Variety stores are the enigma, having no real trend.



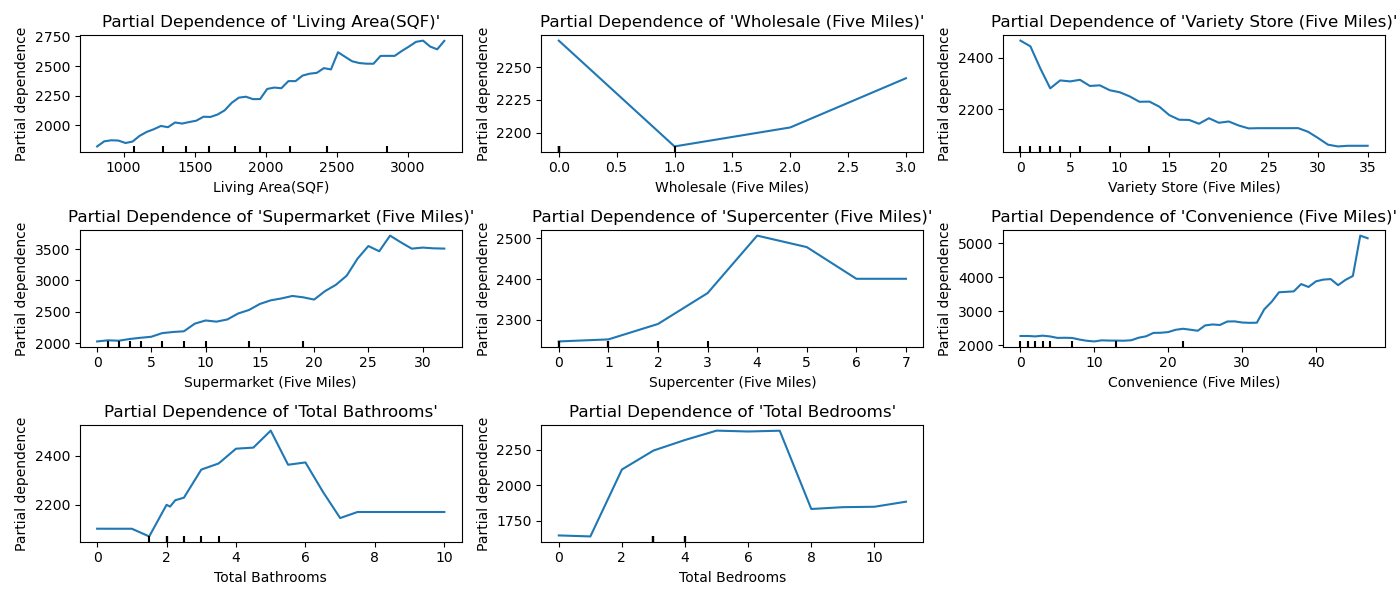
(Fig. 8: XGBoost PDP of Distance Features)

The density PDPs technically show the same results as the distance PDP, just in a different direction because a higher number would mean more value within that certain mile radius. For example, in Fig. 9 we can see that supermarkets have a somewhat steady increase along with predicted outcome for housing prices. In other words, the more supermarkets within one mile, the more likely housing prices will be higher. This was true for every distance feature except variety stores, which we already know is an anomaly. Supercenters do flatten, but that may be because there weren’t enough houses containing four within the one mile distance to have an effect. It is also of note no house had more than one wholesale store within one mile.



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

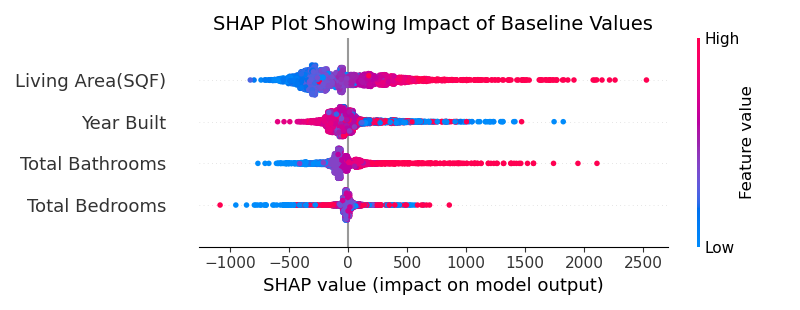
According to Tables 14, the density-one had a tendency to be the worst performing model in comparison to the baseline. Since the distance scores surpassed it in the score comparison, it may be a poor measure of predicted outcome. Luckily, Fig. 10 shows a more concise version of Fig. 9, with a clear increase in the predicted outcome for housing prices as the number of POIs increase. The wholesale feature drops at one, explained by Fig. 9 as no feature had more than one wholesale store at one mile. Interestingly, variety stores also continue their downward trend, possibly giving evidence of a negative externality.



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

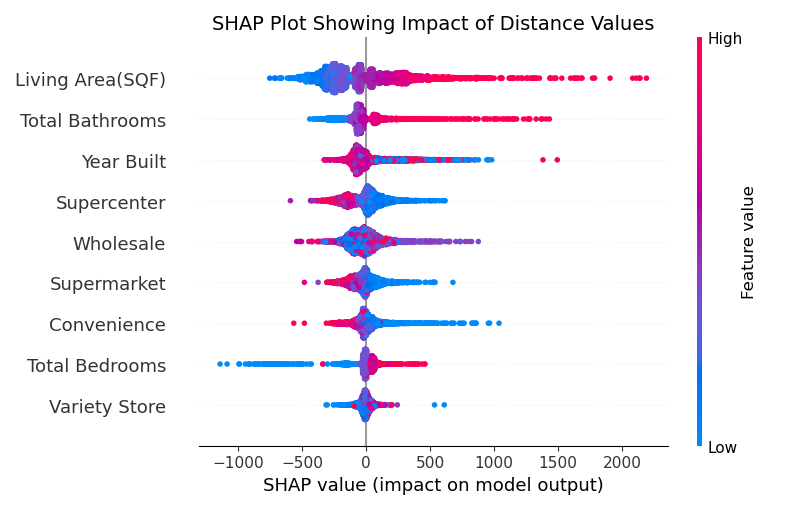
**Impact**

Now that we know how each of the features affect the predicted outcome, we can also show the overall impact of the features using the SHAP (SHapley Additive exPlanations) plot. Essentially, we want to see if a feature changes the target value in a certain direction for every observation reported. Features with a positive value increase the target, and those with a negative value decrease the target. Fig. 11 is the baseline SHAP summary plot, which explains the overall predictions based on the data the model "sees". Each feature is explained differently for the baseline. For example, for living area what it’s saying is that “red” values (housing properties with larger living areas) tend to be on the positive value of the SHAP (which means they contribute positively to predictive housing value). SHAP measures feature importance differently, and therefore ranks the features differently. The ranking reflects the average impact of each feature on predictions, instead of frequency associated with the feature across all tree splits as with the XGBoost.



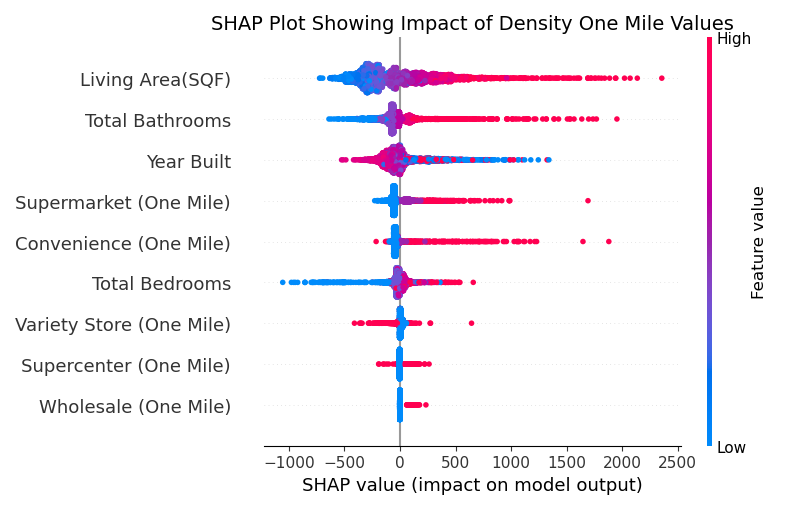
(Fig 11: SHAP Summary Plot of Baseline Features)

When it comes to distance, the baseline features don’t seem to actually be affected by multicollinearity to a significant degree. However, we can see the trend that the PDPs hinted at. When it comes to at least supercenters, supermarkets, and convenience stores, the lower the distance is, the greater the increase in housing value. The opposite also seems to be true, especially for supercenters, where those stores further away seem to have a negative impact on predicted values. However, wholesale and variety stores seem to report mixed results, which could be explained by both their low feature importance and their erratic partial dependency.



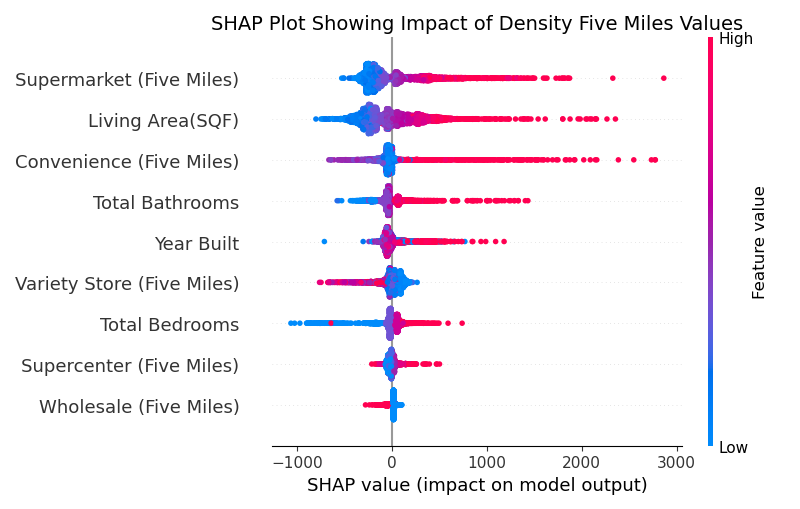
(Fig. 12: SHAP Summary Plot of Distance Features)

Fig. 13 once again shows a reverse in trend for the features, only because a “red” high value would now mean something positive for the impact. We can see that, even at one mile, supermarkets and convenience stores lead to greater predicted housing value, if there are more of them. However, the results are inconclusive for the rest of the stores, even though variety stores do have a number of stores within one mile.



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

If we observe the five miles plot, the impact is more apparent. Fig. 14 shows that supermarkets have become the most important feature, surpassing even living area. Convenience stores keep with the same trend, and supercenters start to show that they very slightly have a positive impact on housing values. However, for the second time, as Fig. 10 demonstrated, the number of variety stores seem to have a negative impact on housing values. Although a concrete determination cannot be made, preliminary results would also indicate that the number of wholesale stores also have a negative impact on the target.



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

**Conclusion**

H0 was that the selected POI category, grocery stores, does not affect housing prices. Table 12 showed us that there is a significant difference between the base model and features that expressed distance between the POI and house locations. Table 1 told us that the distance scores were superior to the base scores (Table 10). Figs. 8 and 12 then expressed the importance and impact that each distance feature has on housing prices, demonstrating that, for the majority of features, being closer to the store leads to an increase in predicted value.

Because H1 is that housing prices are affected by proximity to grocery stores, it was inferred that the number of grocery stores also mattered. This was confirmed by Tables 14 and 15, which demonstrates a significant difference between the density models and the base model. Figs. 9 and 10 also showed a general increase in the predicted outcome for the target as the values of the features increased. Figs. 13 and 14 showed this more clearly, with supermarkets becomes the most important feature at a five mile radius, assuming it is calculated based on average impact instead of frequency or gain.

Overall, it would seem that we can reject the null hypothesis based on prior literature and from what we discovered. The general trend is that, the closer a grocery store is to a house, the greater the predicted housing values. In addition, if there are more grocery stores around the house, then its predicted value may be higher. The only exceptions are variety stores and possibly wholesale stores, which may actually decrease housing value the more stores there are.

**Model Evaluation: Advantages and Disadvantages**

Five main models were used, split into twenty-five because five different sets of data were being used. It was determined that the strongest model was the density-five XGBoost model, and the weakest model, excluding the base, was density-one MLR model. The models that delivered the best scores were consistently the XGBoost models, while the models that delivered the worst scores were always the linear regression models. The Random Forest models had a tendency to be somewhere in between, always being offset from the XGBoost models in a negative way, but still superior to the regression models. XGBoost models are non-linear and can handle complex relationships between the features and target, while MLR is just that, a linear statistical model, assuming linearity, and struggles too much on large datasets.

Even when we tried to fix the MLR model by adding regularization via the Ridge and Lasso, it wasn’t enough to improve the scores to that of the typical XGBoost’s level. The problem with both of those models is that they still assume linearity, with coefficients (based on Maximum likelihood estimation) that would have contradicted the XGBoost results. Just as an example, the Lasso coefficients for distance assumed every distance coefficient was in the negative direction except for variety stores.

This isn’t to say that the XGBoost model is perfect, as it’s often difficult to interpret directly (in our case it required two different graphs), and can be extremely sensitive when it comes to tuning. For example, overfitting can easily occur not from a data issue, but from a slight change in the alpha, depth, or number of estimators. The Random Forest model has similar issues in regards to its parameters. The Random Forest is simpler, at the cost of it being computationally expensive and certain parameters affecting the result drastically, and therefore taking much time. This specific design had a significant amount of data, so XGBoost quickly became favored because it not only answered the research questions, but did so in a timely manner.

The feature importance graphs were the easiest and quickest ways to display the results desired. They are also a direct result of what a Random Forest would typically try to convey. While on the surface the results were interesting, they technically didn’t answer the research question. Fig. 3 said wholesale was the only store type that was more important than a base feature. Without any other context, what would this mean? If anything, this would say we cannot reject the null hypothesis as the other stores seem to be passing into irrelevancy. Luckily, the density models showed us that this was not the case, but the point remained that feature importance needed something else to support it.

The PDP graphs were probably the most substantial in explaining how the features affected the target, especially since each “graph” was technically independent of each other. It is straight to the point: is there an increase or decrease of the predicted average housing prices along with the feature? However, the problem with a PDP is just that, they ignore feature interactions entirely and can lead to misleading results. “Partial dependence” is also an average value, which has a tendency to “smooth” data and ignore nuances. For example, it’s likely the variety store results in Fig. 8 probably look more like “noise” than what is displayed.

Finally, we used the SHAP as an attempt to visualize actual value scores and how they impacted the target. The problem with the SHAP display right away is that it’s difficult to interpret, with many of the values hovering around zero because of our data. This resulted in the plot looking like frequencies than something that could be interpreted. Still, if one can see past the noise, the SHAP is one of the more beneficial plots. It provides some of the most information out of all the plots used, and without it we wouldn’t have been able to confidently seen how the POIs interacted with the target.

**Further Research**

Atlanta, Georgia was chosen because prior research was done in the area in regards to POIs and housing data. It is considered an "easy pick" for studies due to being a major economic center in the Southeastern US, along with being a host to major public transportation. With the results found, it is highly recommended, even encouraged, to either individually explore other regions of the US using the same or similar methods or compare these results with results from other regions. This would allow us to see if it was actually the demographics and economic composition of Georgia that affected the results, or if grocery stores really do affect housing prices in a similar way universally.

It is also advised to either expand or focus in on one of the features used in the design. The features used were chosen so from of a mixture of both Overpass’s limitations and from other designs in the literature. The way that these features are created can significantly, if not fundamentally, change the entire design of the study. For example, some scholars might not consider supercenters to be a viable feature, and would instead group brands like Walmart and Target into the supermarket category. Others might consider large chains like Publix to be a supercenter. Further, some might prefer to break up stores into groups, adding regional and national chains as features instead of just having a generic “supermarket”.

We picked out density and distance and compared them against the baseline, but the literature hinted at other features that also influence how customers shop. Having these other features, like store quality, price, assortment, and “branding” be paired with the proximity/density factors would be interesting to see how they interact with each other. Maybe some individuals would be willing to ignore distance for certain stores if the quality was high, thereby increasing the housing prices in an unexpected way.

Furthermore, we were examining the entire Atlanta, Georgia metropolitan area, which includes the rural countryside. Batt and Chamhuri suggested that there was a difference in what kind of income “central locations” stores attracted, where low income consumers were essentially forced to shop locally because they didn’t have the means to travel. Our study did not take this kind of disparity into account, combining both urban and rural data. In other words, the way a grocery store’s proximity in a central city than in a rural region may drastically not only be different, but also have very different effects on housing prices.

As for the models themselves, three things can be noted: the R2 scores were lower than .90, we put too much comfort in one type of model, and there was at least some overfitting, even if it was intentionally allowed. In a perfect scenario, we would want a design that has models competing against each other where certain ones have similar scores, rather than all the models having statistically different scores. One suggestion could be to use a Kernel Ridge regression as a supplement for the Ridge regression. However, since XGBoost was so effective, we could try Stochastic Gradient Boosting, which adds randomness to improve generalization. A hybrid model could be attempted like a Voting Regressor, which aggregates predictions from several models using weighted averaging. Finally, instead of an XGBoost we could try a specialized gradient boosting, such as an NGBoost, which models uncertainty by estimating probability instead of points.

**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 |

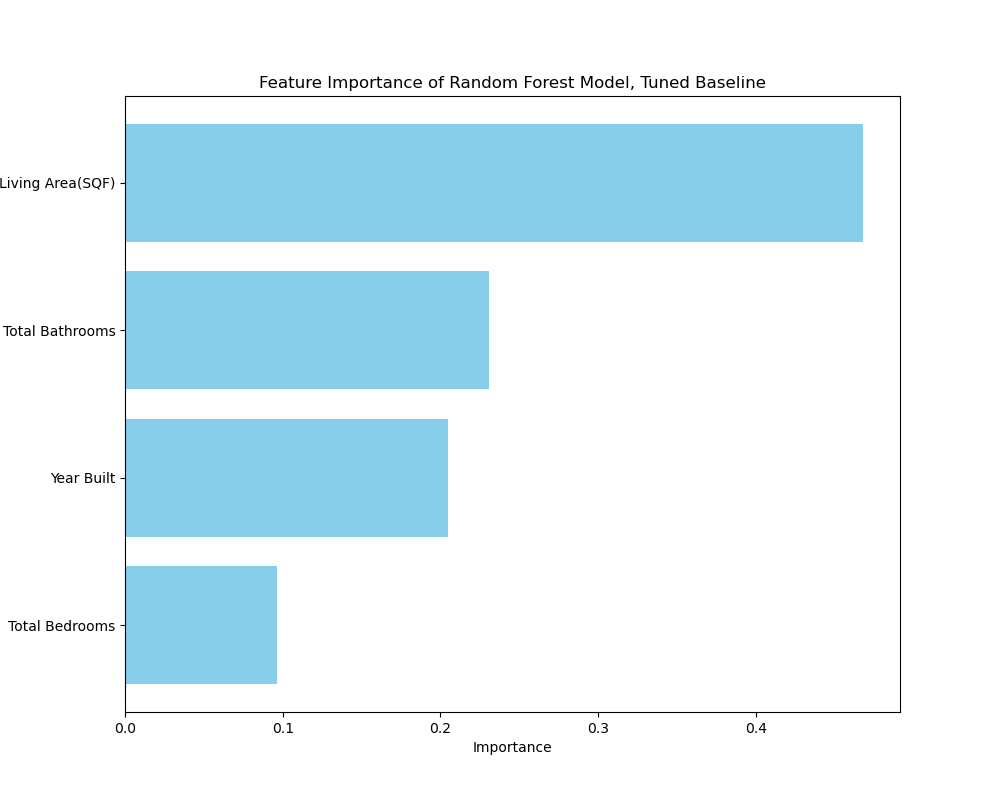
Table A4: Score Comparison for Random Forest Three Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

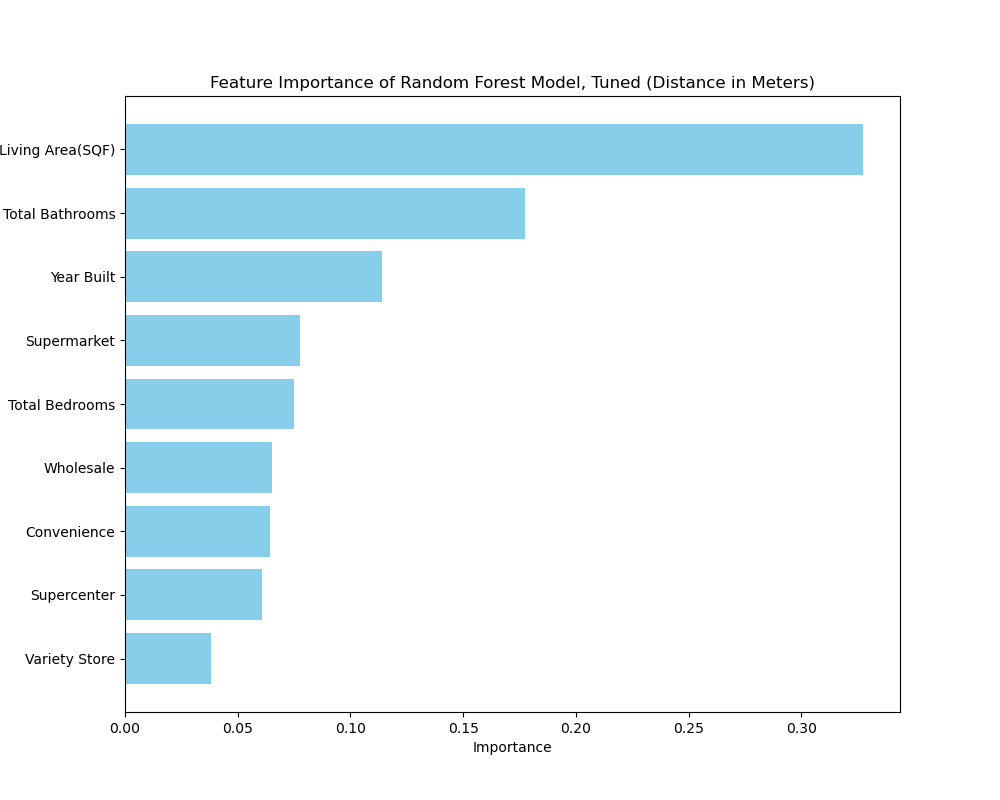
Table A5: Score Comparison for Random Forest Five Miles

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

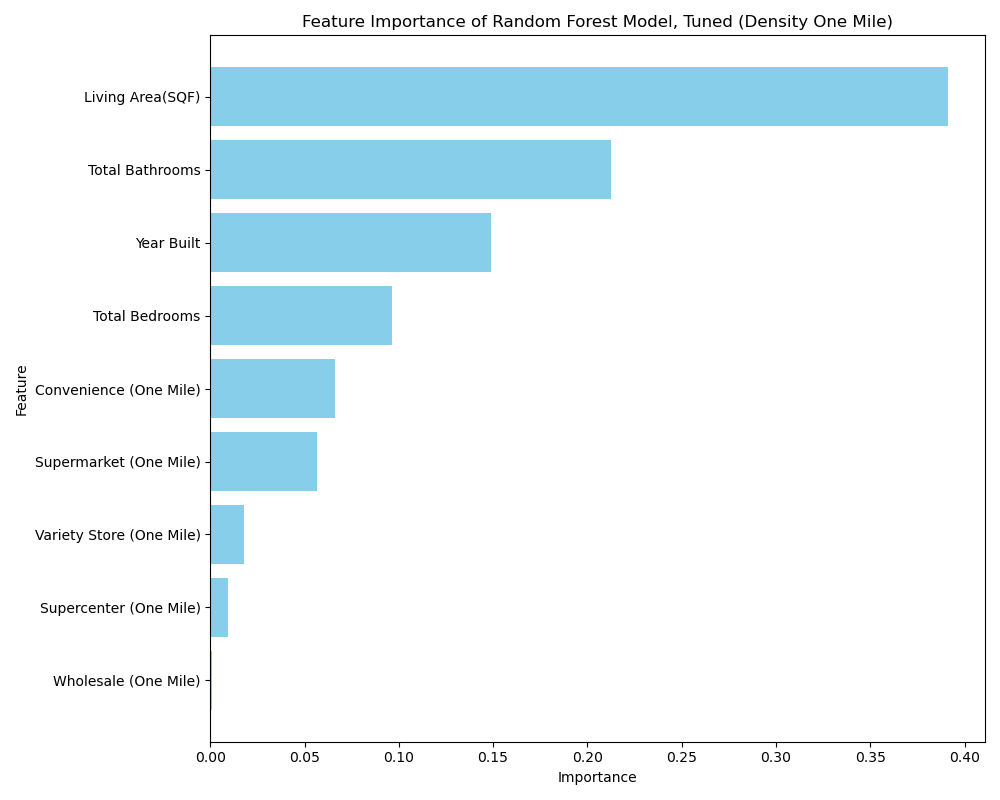
**Group B: Random Forest Feature Importance**

****

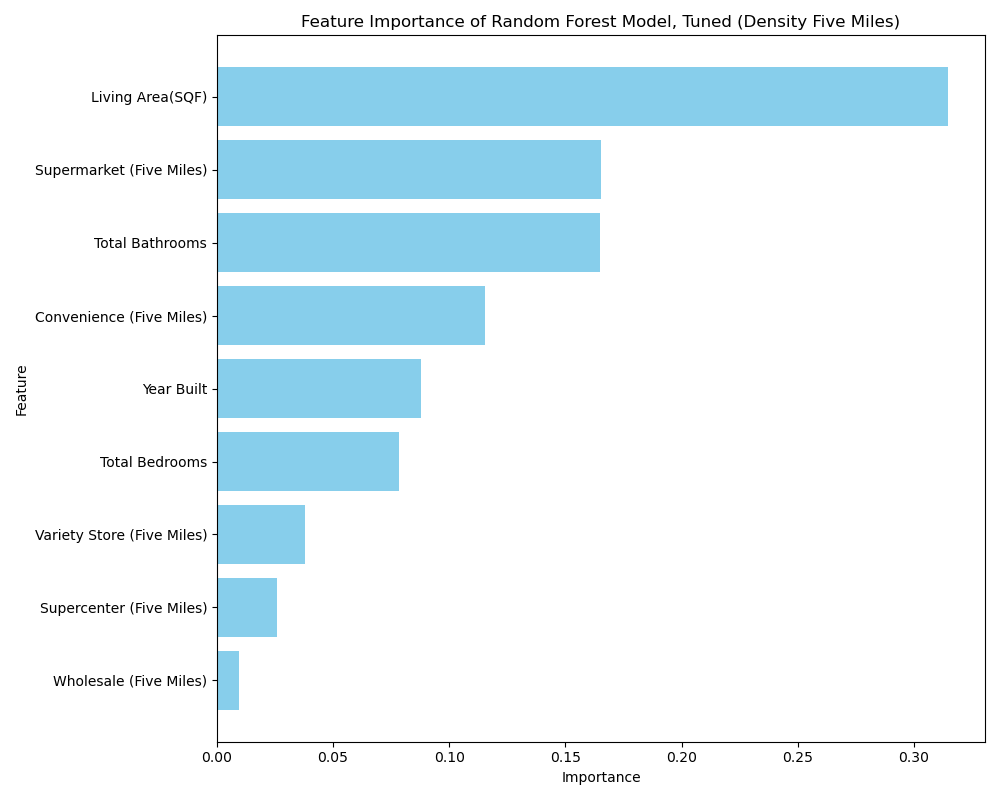
(Fig. B1: Random Forest of Baseline Feature Importance)

****

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

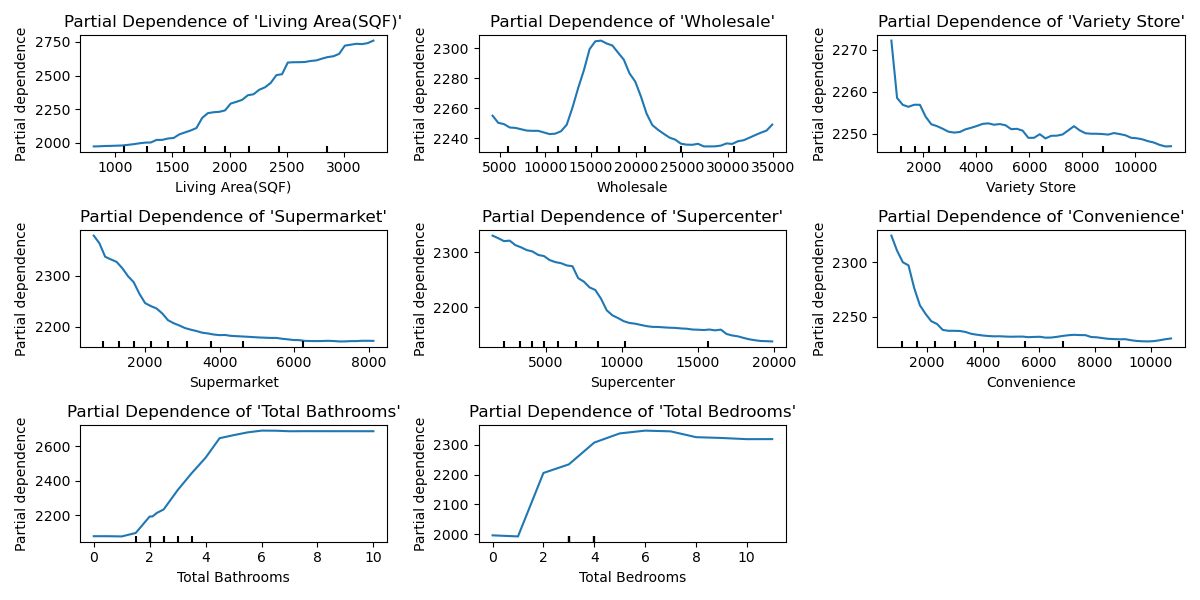


(Fig. B3: Density One Mile Random Forest Model Feature Importance)

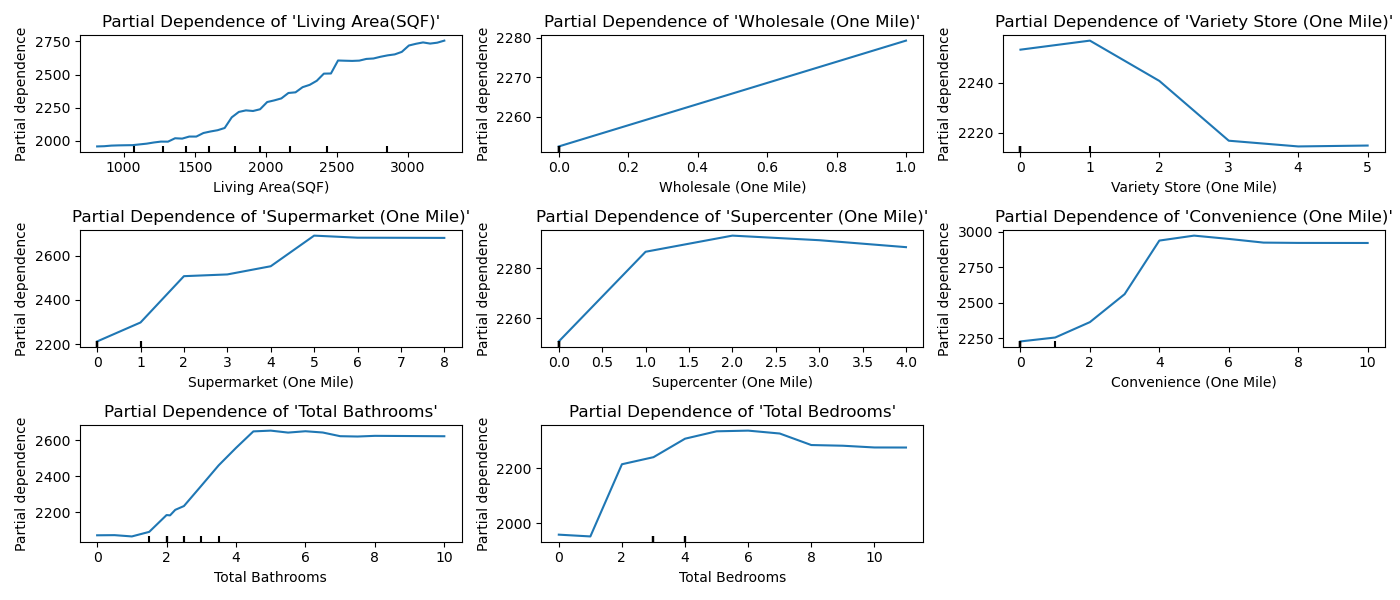


(Fig. B4: Density Five Miles Random Forest Model Feature Importance)

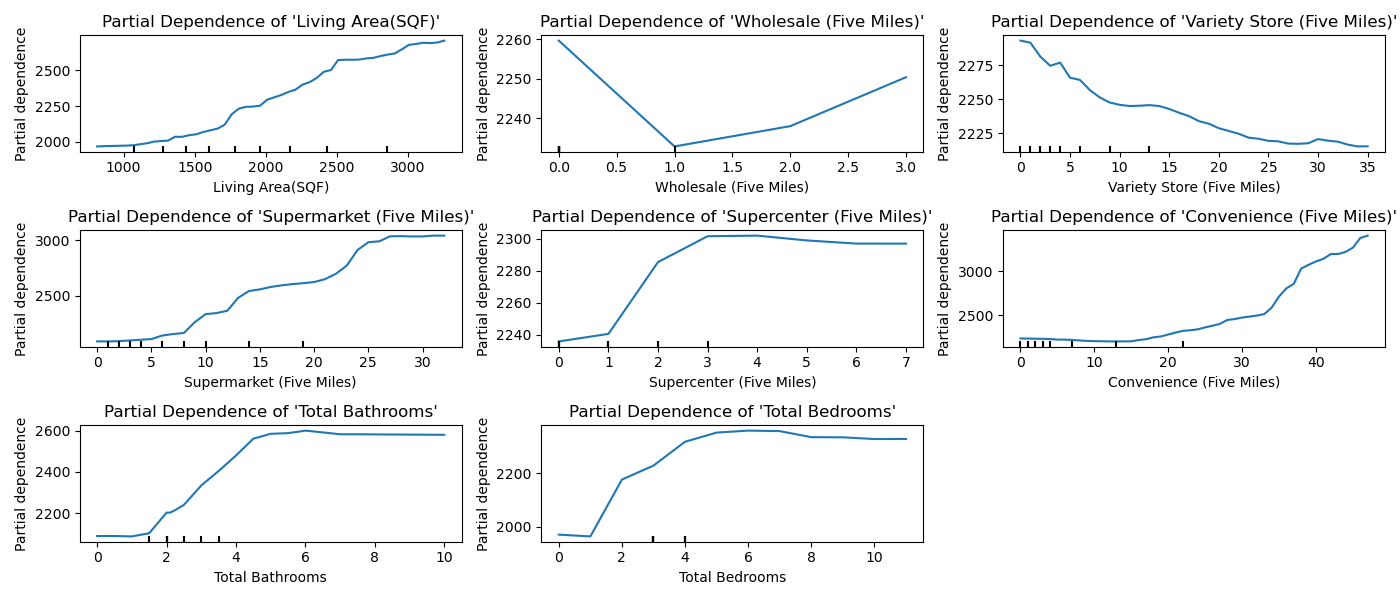
**Group C: Partial Dependence Plot**

****

(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 |

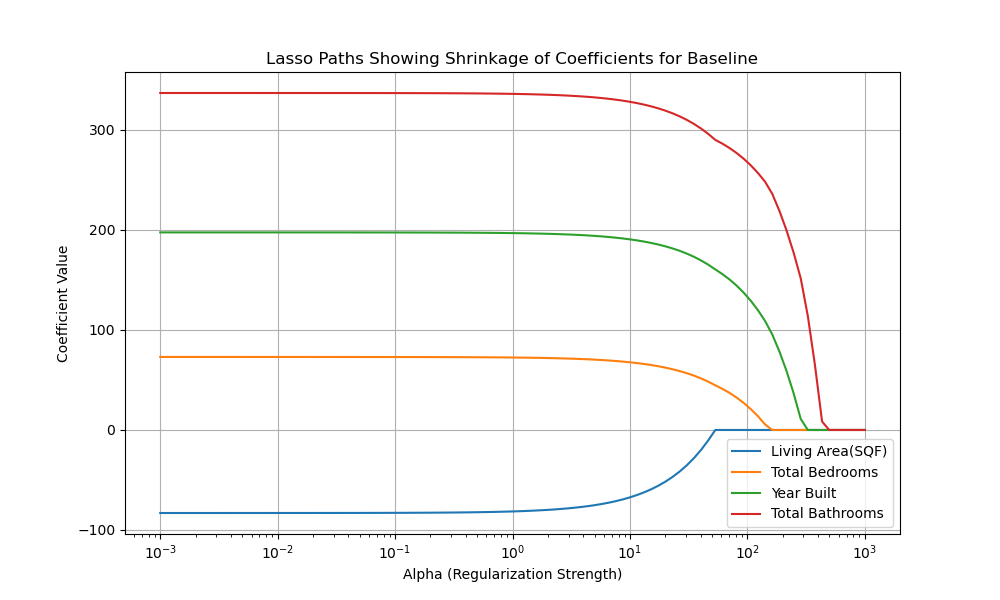
Table E4: Score Comparison for Ridge Three Miles (Log)

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

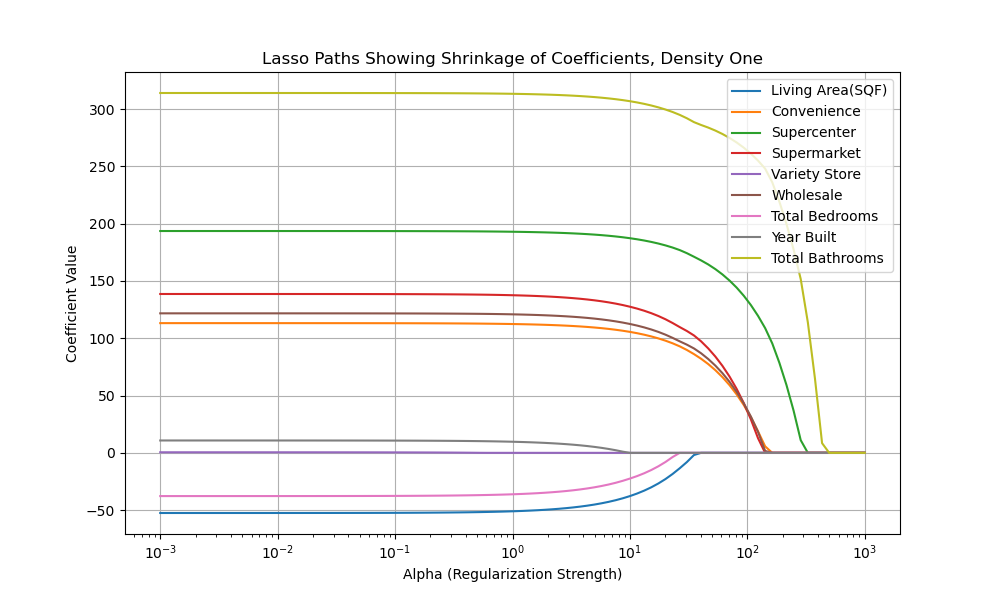
Table E5: Score Comparison for Ridge Five Miles (Log)

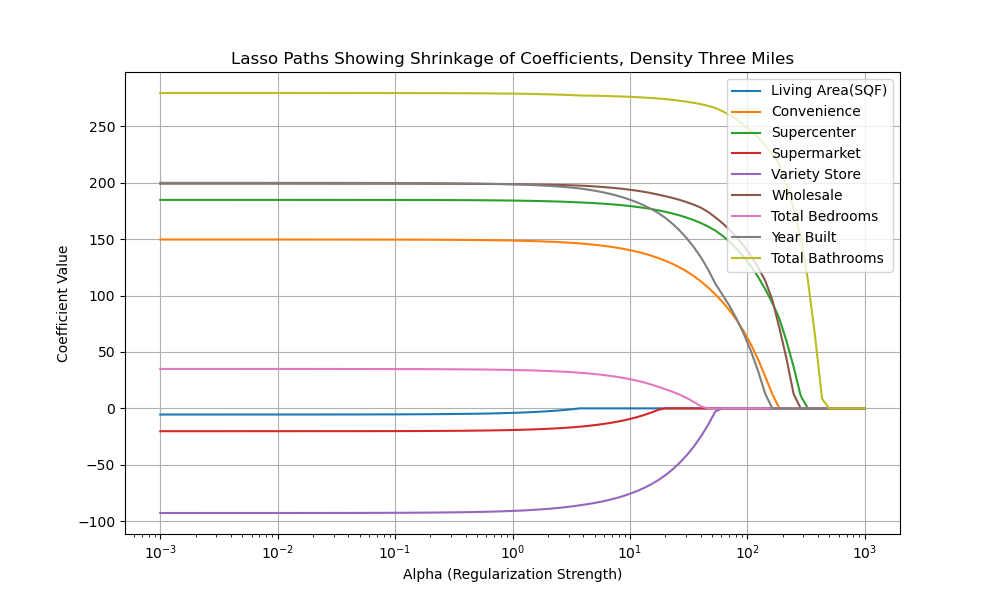
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Test Set** | **Train-Val Set** | **Validation Set** |
| **R2** | 0.4681 | 0.4572 | 0.4894 |
| **MSE** | 0.0588 | 0.0596 | 0.0599 |
| **RMSE** | 0.2426 | 0.2442 | 0.2448 |
| **MAPE** | 0.0211 | 0.0216 | 0.0219 |

**Group F: Lasso Regularization Graphs**

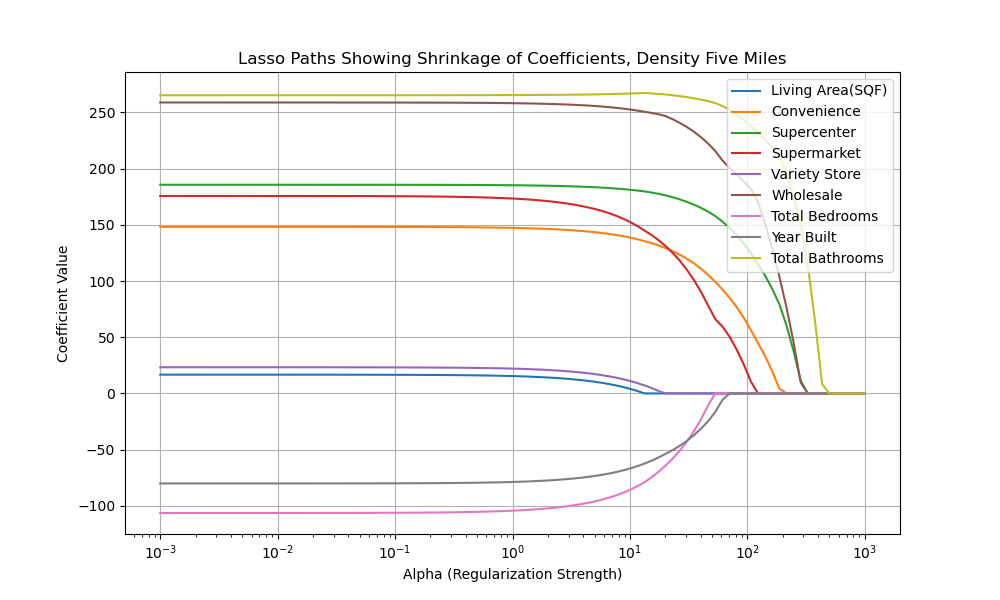
****

(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.