Lauren Flemmer
STAT222
Team C

Zillow Price Prediction

1. Introduction

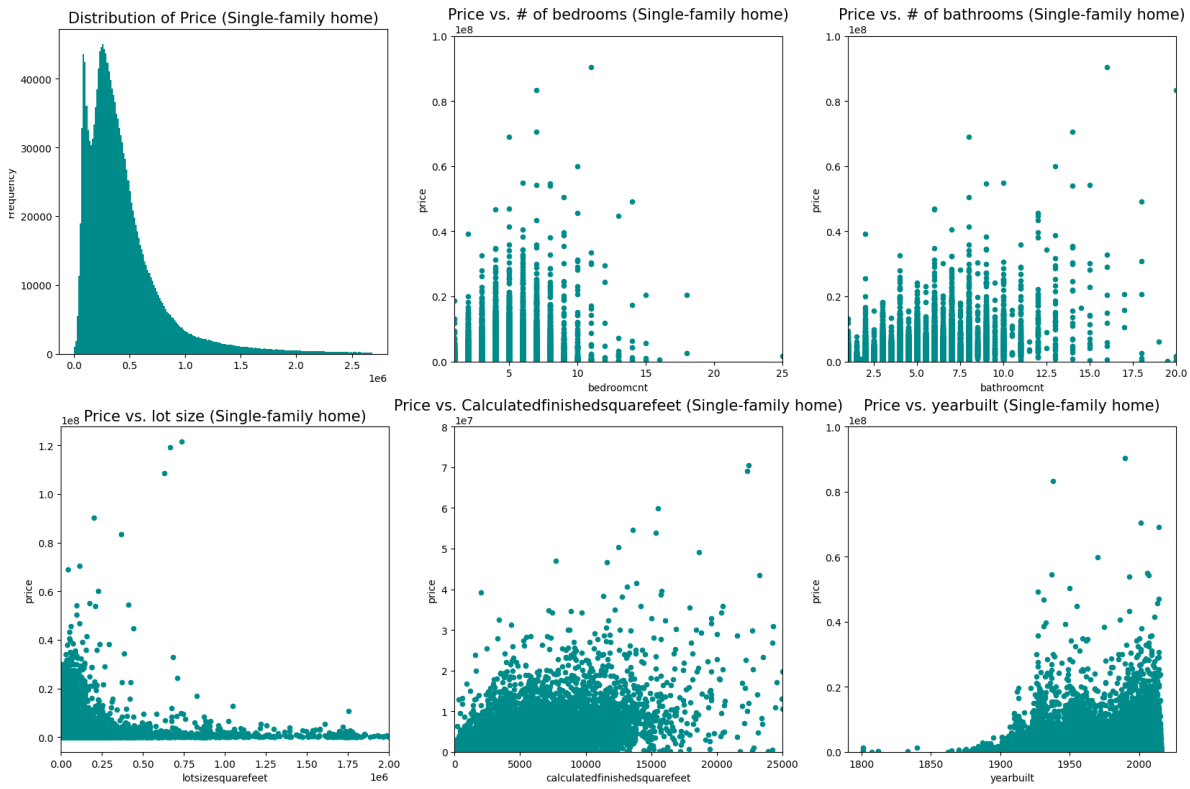
In recent years, pricing models have begun to play a large role in commerce, specifically real estate. With the rise of online real estate marketplaces, such as Zillow and Redfin, the concept of algorithmically pricing homes and other properties has become a primary element of these companies' products. By developing pricing models, like Zillow's *Zestimate,* for properties listed on the sites, these companies are generating meaningful information about those properties that will inevitably inform the decisions and thinking of buyers, sellers, lenders, and real estate agents, so companies like Zillow play an immense role in the real estate market, even involving properties that aren't necessarily listed on the site. Besides buyers and sellers, metrics like Zestimate are used by parties like appraisers, insurance companies, and mortgage lenders, to help inform similar decisions in the buying or selling processes. Due to their increasing popularity and influence in the real estate industry, it becomes more and more important to improve these pricing models to ensure that they are accurate. Because so many aspects of the selling or buying processes involve pricing models like Zestimate, having an accurate model is crucial to keeping the real estate market operating smoothly. My team recreated Zillow's *Zestimate* algorithm in order to accurately predict the value of homes in the Los Angeles, Ventura, and Orange Counties.

The data we used came directly from Zillow, from a Kaggle competition they held starting in May 2017. The goal of the competition was to build a pricing model comparable or better than Zestimate, for homes listed on the site in 2016. Since the data was collected and used by Zillow for the purpose of calculating Zestimates, it is ideal for attempting to replicate and improve the existing Zestimate model. However, instead of using the log error as our variable of interest, we used the property's assessed value, a proxy for the price, as our response. Our GitHub repository can be found here.

2. Exploratory Data Analysis

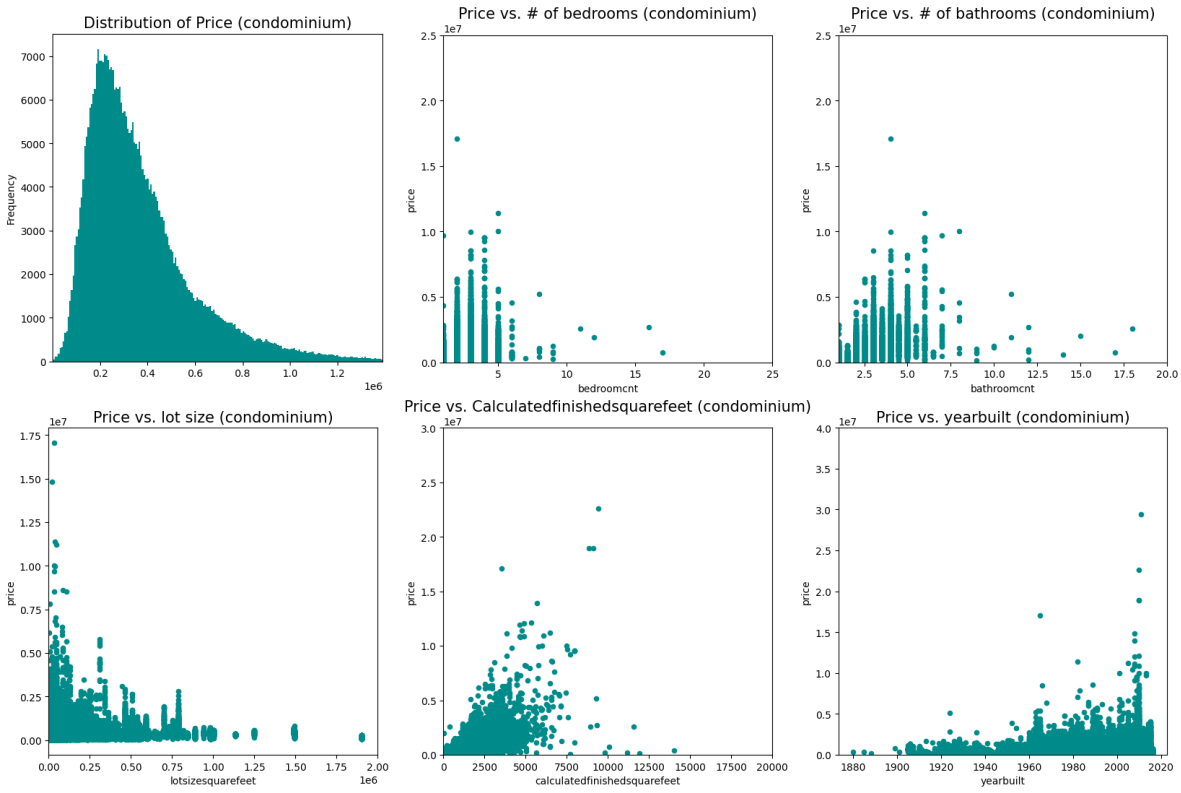
When initially conducting EDA on our features, we noticed that there were some relationships that didn't quite make sense to us. When we looked deeper into these relationships, we realized that attributes of different types of properties had different relationships with the value of the property. For example, the relationship between lot size and home value will not be the same for a mobile home as it is for a single-family home, since purchasing a single-family home includes purchasing the land it is on, which is not the same case with mobile homes. Therefore, different types of properties should be separated into different categories before performing any EDA. The "propertylandusetypeid" feature details 25 different types of properties, such as "Single family home", "Duplex", "Triplex", "Condominium", "Mobile home", etc. For many of these categories, only a small percentage of our data belonged to that particular category, so we decided to focus only on the 4 main property categories: Single family homes, making up ~72% of the total properties, Condominiums, making up ~16% of the total properties, Multi-unit homes (consisting of Duplexes, Triplexes, and Quadruplexes), making up ~6% of the total properties, Townhomes, making up ~2% of the total properties, and Mobile homes, which make up ~1.6% of the total properties. Below is the EDA for each of these property categories, showing the relationships between various features and the price of the property.

Figure 1: EDA for Single-family homes



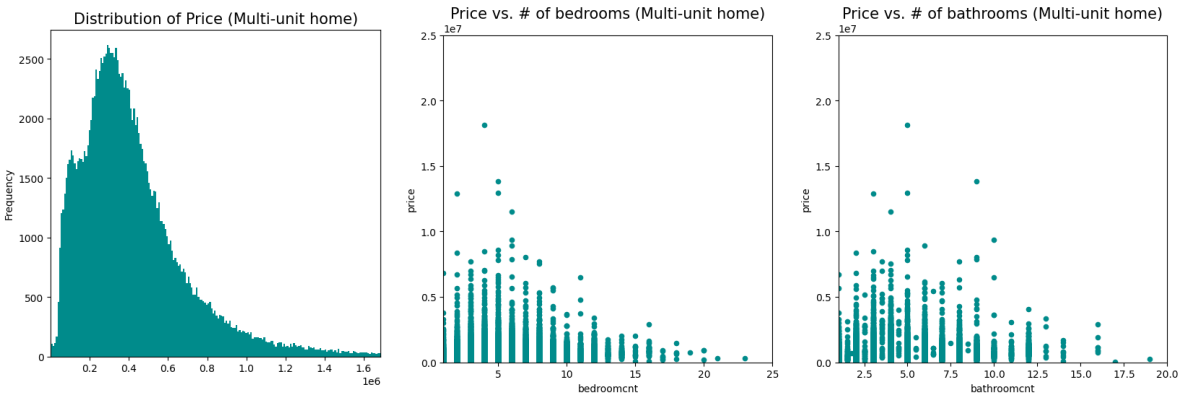
From observing the plots comparing the attributes of single-family properties with their value, we can see some relationships that intuitively make sense- for example, for the number of bedrooms, number of bathrooms, total square footage, and year built, we see that there is a positive linear relationship between these features and the value of the property. This is expected, since as a house becomes newer and larger, whether that be by number of rooms or square footage, we expect it to be more expensive. An interesting relationship here is the relationship between the lot size and the value, which seems to have a negative trend, i.e, as the lot size of a single-family property increases, the value decreases. Thinking about this, we would expect a property on a larger piece of land to be more expensive, but considering these homes are in the Los Angeles area, this observed relationship begins to make more sense. Near more populated metropolitan areas, for example Downtown LA, there are few properties with large lots, but the prices of these properties are very high due to their location. However, as we move farther away from these expensive metropolitan areas, we tend to move out into the suburbs, or more rural areas, where there is more space, i.e, larger lot sizes, and real estate/land is generally cheaper. Although a confusing trend at first, it becomes clear when considering the background and setting of our data set.

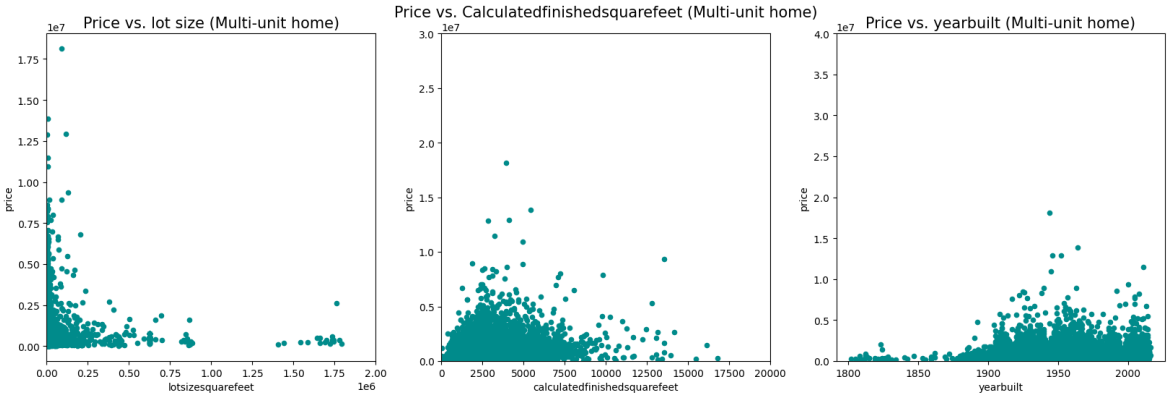
Figure 2: EDA for Condominiums



Similar to the case of single-family homes, we can observe some obvious trends for condominiums. Again, the number of bedrooms, number of bathrooms, total square footage, and year built have a positive linear relationship with the value of the property. As we saw with the single-family case, there is also the same interesting relationship between lot size and value, where we see decreasing value as the lot size increases. Although the lot sizes of condos and single-family homes can't be interpreted exactly the same, since a condo might be in a larger building, comprised of multiple condos (where the lot size likely represents the lot size of the entire building, rather than the lot size of the individual condo), it would make sense that this relationship would have the same explanation as single-family homes. In general, a condo's lot size (whether that lot size refers to the individual condo or the larger structure the condo belongs to) will be larger in areas that are less populated and have cheaper land.

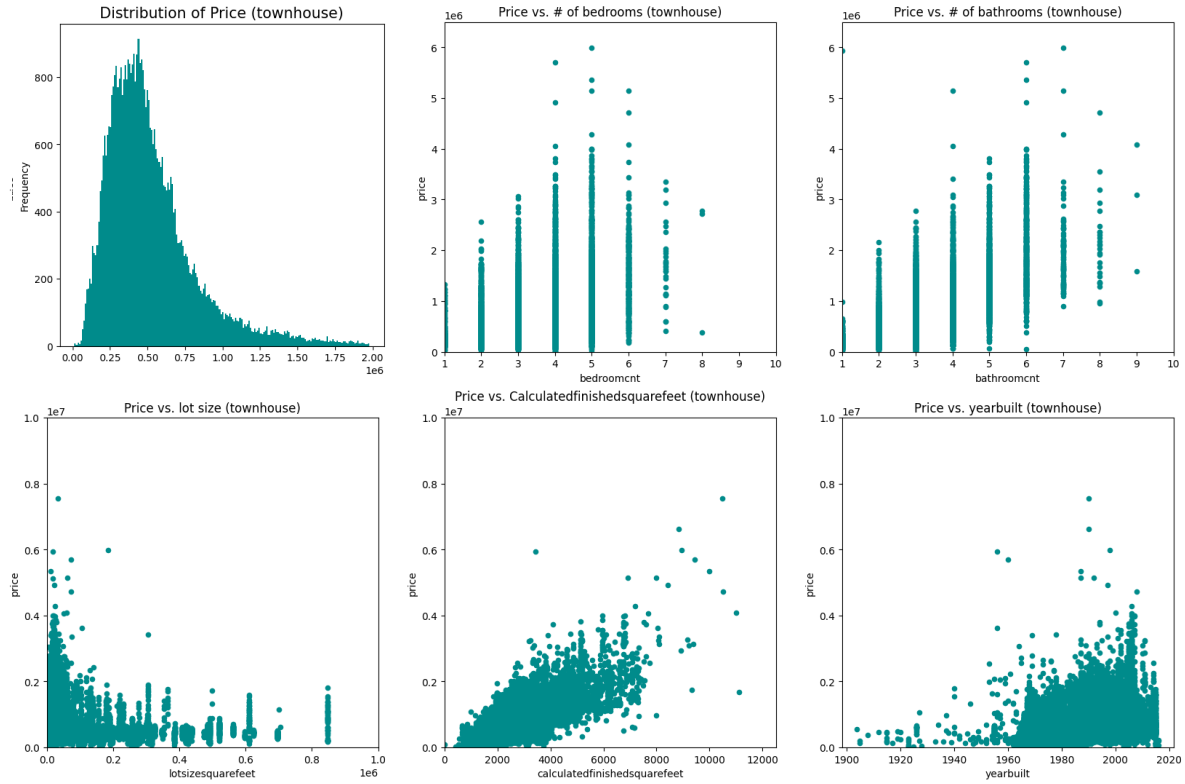
Figure 3: EDA for Multi-unit Homes





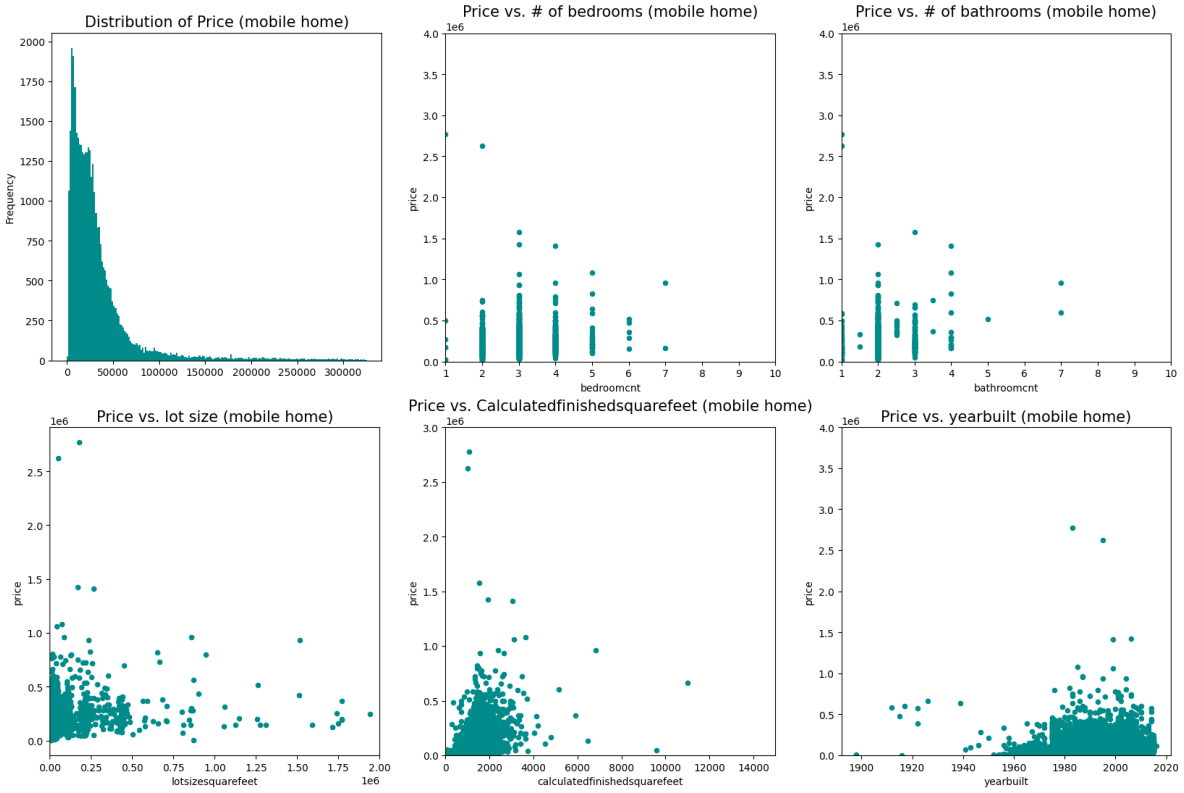
Now looking at multi-unit homes, we observe something very interesting. The relationship between the number of bedrooms, number of bathrooms, and square footage begin to have a negative relationship with the value. This doesn't make sense at first, but when looking deeper into the plots, it becomes clearer why we are seeing this relationship. The value begins to decrease for very large number of bedrooms, bathrooms, and square footage. For example, for square footage, the value begins to decrease when we hit ~ 7500 sqft. Considering these homes are in the Los Angeles area, a lot of the larger homes are likely in less populated areas, where there is more land. Since these areas are less populated, they are cheaper to live in, which might explain the low property values. This idea extends to the number of bedrooms and bathrooms. Imagine a home with 20 bedrooms- it is likely in the suburbs or in an area far from Downtown Los Angeles, meaning it is also cheaper than if it were downtown.

Figure 4: EDA for Townhouses



For townhouses, the trends we see are as expected- features like number of bedrooms, number of bathrooms, square footage, and year built have a positive linear relationship with the value of the property. Again, we see that for large lot sizes, we see a decrease in the value of the property, which can be explained again by the fact that these properties are likely in areas farther away from a highly populated metropolitan area, where land is cheaper, allowing for a property to have a lower value.

Figure 5: EDA for Mobile Homes



For mobile homes, we observe similar relationships between relevant property attributes and the value of the property, but on a smaller scale since there are fewer mobile homes in the dataset, and their features overall have less extreme values.

3. Dealing with Missing Data/Feature Selection

One of the biggest challenges with this dataset is the amount of missing data, and determining how to handle those missing values. For many of the features in our dataset, there was too much of the data missing to extract any meaning from it. For example, many of our features had over 90% of their values missing. In this case, it doesn't make any sense to do any sort of imputation, so these variables had to be dropped. We imputed missing data using 3 different methods: zero imputation, mode imputation, and median imputation.

We imputed missing values with zero when the implied value of the 'NA' is zero. For example, in a variable like "garagecarcnt", the number of car garages the property has, if a property has an 'NA' car garage, it is implied that it does not have a garage, so it is reasonable to replace those missing values with zero. We did this for all features where it made sense. If most of the observations of a particular feature

had the same value, it was reasonable to replace any missing values with the mode of the feature. For example, with the feature “airconditioningtypeid”, which is a categorical variable representing the type of cooling system present in a property, the majority of the properties had Central AC, so it was reasonable for us to impute missing values with the mode, replacing these ‘NA’ values with the ID corresponding to Central AC. We did this for all variables where one particular value or category dominated. Lastly, for features where the distribution of values was not as uniform, we used median imputation to impute missing values.

4. Modeling

For predicting the value of a home, we trained multiple machine learning models in order to find the one that performed best. We fit linear models, as well as tree-based models such as XGBoost and LightGBM. Because the different property types have varying distributions for their different attributes, they have different pricing standards and, so we decided to train separate models for each property type, as well as on the full data. By doing this, we would be able to compare model performance for different property types, and determine our predictions for specific property types are more accurate than our predictions on the data with all property types combined. Additionally, for our tree-based models, since they are able to handle missing values we trained them using our full imputed data, with the imputations discussed in Section 3, as well as on the full unimputed data, to determine if our imputation methods made a difference in model performance. Through modeling the value of homes in the Los Angeles area, we hope to not only find a way to accurately predict the value of a home, but also to determine which features were most important to a home’s value by investigating the feature importance for each property type. Since we were predicting the assessed value of the property, it was important for us to remove all tax related information from the final dataset before training our models, since the assessed value is based on property taxes. Additionally, we performed other feature selection methods, such as removing any redundant or highly correlated variables. Results from our models will be discussed in the next section.

5. Model Performance

Since the relationship between our features and price was not exactly linear, some of the assumptions for linear regression were not met. Therefore, I will only be discussing the results from our other models, XGBoost and LightGBM. For interpretability, we used MAE (Mean Absolute Error) as our main evaluation metric, since it can easily be interpreted as the average difference (in dollars) between our predictions and the actual assessed value of the property. We also computed our “log-error”, defined as logerror = log(predicted) - log(actual), which we used to compare our models’ performance to Zillow’s. Zillow’s “log-error” is defined as log-error = log(Zestimate) - log(Sale Price). Additionally, when we trained the tree-based models on both the imputed data and unimputed data, the performance was very similar, so the results below will be for the models trained on the unimputed data. First I will discuss the performance on the full data (not separated by property type).

Model	**Mean Absolute Error**	**Mean Squared Error**	**Log Error**
XGBoost	179039.20	532766.87	0.15364
LightGBM	184478.94	699186.38	0.20911

From looking at these results, we can see that the XGBoost and LightGBM models have very similar MAE, with the XGBoost model performing slightly better. Interpreting this MAE in terms of the problem at hand, we can say that these models are off, on average, by around $180,000 in their predictions. Now I will discuss the model performance when the unimputed data is separated by property type.

Property Type	**Model**	**Mean Absolute Error**	**Mean Squared Error**	**Log Error**
Single-Family Home	XGBoost	197485.359375	450473.8125	0.1850444354378
	LightGBM	195330.3409371	451593.392419	0.508649075082
Condo	XGBoost	77634.57	113657.97	0.06258
	LightGBM	99871.84	186777.76	0.08083
Multi-Unit Home	XGBoost	191020.53125	311625.625	0.19889733
	LightGBM	190575.962703	302244.788110	0.2383195815
Townhouse	XGBoost	119774.8515625	245363.84375	0.1778212095
	LightGBM	121519.206390	188367.633073	0.19122900368
Mobile Home	XGBoost	22409.2246093	50282.3515625	0.28762963
	LightGBM	22074.8708110	45081.7199212	0.3507018692

We can see that when the data is separated by property type, some of the models perform better than they did on the full data, and some perform slightly worse. The models for mobile homes had the best performance, having an average error of around $22,000, while the models for single-family homes had an average error of around $195,000. However, all of the models for the separated data had better performance than the models on the full data. This illustrates how the differing relationships among features for different property types are better learned when there is a separate model for each type. Overall, our models performed fairly well, but not as well as we had hoped. The value that lies in these models does not come from how well they predict the value of a home- the insights we gathered from these models came mostly in the form of feature importance, which will be discussed in the next section.

6. Feature Importance

As mentioned before, the most informative takeaways from our project came in the form of feature importance. Using the feature importance scores from our XGBoost and LightGBM models, we found the top five most important features for each model. Interestingly, all of our models, for both the full data and the data split by property type, had the same five most important features (some in different orders). The five most important features in predicting the assessed value of a home were the finished square feet of the home, the distance to the ocean (in miles), the distance to Downtown Los Angeles (in miles), the lot size (in square feet), and the year the home was built. Looking at this result, all of these features make sense- before starting this project, we all had some general idea of what was the most

important to the value of a home. For example, we knew that newer homes would likely be more expensive than older homes, and that the larger a home was, the more expensive it would be. However, we didn't think that the variables we created, the distance to the ocean and the distance to Downtown LA, would be this important in predicting the home value. Overall, our findings suggest that factors such as location and proximity to important locations can play a significant role in determining the value of a home, and that machine learning models can effectively capture these complex relationships between features and home values.

7. Conclusion

In our project, it's worth noting that we didn't have access to the actual sale prices of the properties, which is what Zillow's *Zestimate* predicts, and how their log-error is calculated. Therefore, we were unable to directly compare the performance of our models and Zillow's, since our focus was on predicting the assessed value of the properties. In that regard, we were able to achieve a fairly good level of accuracy. While our models could have performed better, the true value of our project lies in the insight we gained from determining the feature importance of our models. By understanding the factors that significantly influence the value of a property, we were able to gain a much broader understanding of how the real estate market operates. By determining the feature importance from our XGBoost and LightGBM models, we were able to identify the features that played a crucial role in predicting the assessed value of a property. The fact that our models were consistent in identifying the same five most important features for predicting the property value across different property types provides further credibility to our results. This suggests that these property attributes hold this importance across a wide range of properties and are not limited to a specific category. In addition, the prominence of certain features in our analysis of the feature importance was surprising. In conclusion, while we couldn't directly compare our model's performance to that of Zillow's *Zestimate*, our project provided us with valuable insights into the real estate market, and what factors contribute to a property's value.