Northeastern US Home Price Prediction

MGT-6203 Group Project Final Report——Team 4

Zilin Ma(zma338),
Wenhui Ma(wma98),
Xin Ye (xye85),
Biyao Zhou (bzhou301)

Github Link: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-4

Table of content

Abstract

The U.S. real estate market plays a pivotal role in the national economy, reflecting economic well-being and progress. Housing, central to this market, serves as a vital indicator of financial stability for individuals and communities. Leveraging the USA Real Estate Dataset from Kaggle, we developed housing price prediction models to discern the primary predictors of housing prices in Northeastern US and investigate key factors influencing the real estate market in the area. We compared several popular modeling methods including multiple linear regression, random forests, SVR, and LightGBM for analysis. The analysis indicated that the LightGBM model family demonstrates a high degree of accuracy in predicting real estate prices. Our duo-model strategy will provide precise price predictions and reveal price disparities across Northeastern States simultaneously. Through this research, we aim to enhance our understanding of real estate market dynamics, providing valuable guidance for all stakeholders.

Introduction

The U.S. real estate market is a multifaceted and pivotal component of the national economy, closely linked to economic well-being and progress. Housing, as the core of this market, is not only an important standard for measuring the financial stability of individuals and communities but also a powerful economic indicator for economists and policymakers to make forecasts and policies. For homebuyers, renters, investors, real estate developers, policymakers, and observers of economic trends, understanding the complexities of housing dynamics is crucial for making wise decisions. This includes grasping housing trends, interpreting price fluctuations, and considering the attributes of specific locations.

Analysis of the housing market helps to increase market transparency, promoting fairness and efficiency, benefiting industry professionals and the public alike. Policymakers can draw valuable insights from it to address issues like affordability. Increases in housing market activity and rising house prices are often signals of economic growth, indicating that people have increasing income and an optimistic economic outlook. At the same time, the state of the housing market can exert pressure on interest rate changes, potentially indirectly affecting the decision-making and monetary policies of central banks. House prices in different regions of a country also reflect the income levels and affordability in those areas, greatly influencing private sector investments and local government decisions on infrastructure spending and fiscal policies (taxation). By observing housing market trends in different areas, we can predict and compare the housing markets of each region, noting whether past trends indicate sustainable, stable economic growth or potential overvaluation in the housing market, leading to an economic bubble. In such cases, monetary policy might require public intervention, and a more conservative investment strategy may be more suitable for private investors.

Objective Statement

The objective of our research is focused on developing a predictive model for estimating the values of properties in Northeastern US without listed sale prices. Central to this endeavor is the evaluation of various models for their accuracy and reliability. In parallel, the research delves into analyzing housing price trends and geographic variations, employing data visualization to track how housing prices have evolved over time in different regions and determining if certain areas exhibit distinct price growth patterns. A key aspect of this analysis involves examining the impact of housing characteristics like the number of bedrooms, bathrooms, and overall house size on property prices, using correlation assessments to quantify these relationships.

Business Impact

The overarching aim of this analysis is to provide comprehensive insights into the trends and characteristics of the U.S. real estate market, catering to the needs of potential homebuyers, investors, developers, and policymakers. This entails focusing on critical variables such as changes in housing features and the influence of geographic location on housing prices, with the intention of uncovering the key factors that drive market dynamics. The potential business impact of this research is significant, offering deeper market insights that can inform more strategic investment and development decisions. For example, investors can utilize this data to identify potential investment opportunities and assess risks, while developers can plan new projects based on a nuanced understanding of housing features and geographic location.

While the model is constructed using housing data specific to the Northeastern region, its internal methodology holds general applicability to other locations, contingent upon the availability of analogous data. The transferability of the model's methodology is contingent upon the existence of comparable datasets from diverse geographical regions, thereby enabling the adaptation and application of the analytical framework to different locales.

Questions and initial hypothesis

1. Price Trends and Geographic Variation: We expect to identify trends in housing prices over time and geographic regions. Our approach will involve visualizing price trends and exploring whether certain areas experience higher or lower price growth rates.

2. Impact of Housing Characteristics: We hypothesize that factors such as the number of bedrooms, bathrooms, and house size will significantly influence housing prices. Our analysis will include correlation assessments to quantify these relationships.

Ultimately, our approach will lead us to determine the final conclusion of our analysis by considering factors such as model performance, price and regional trends, and housing characteristics. While our hypotheses guide our analysis, the goal is to provide stakeholders with a comprehensive understanding of the key factors and trends shaping the Northeast real estate market, allowing them to make informed decisions. The accuracy of our conclusions will depend on the quality and representativeness of the dataset and the effectiveness of the chosen models and methodologies.

Overall of Data

The data was downloaded from Kaggle.com. The original source was from realtor.com. Here are the original dataset feature list:

- bed (# of beds)
- bath (# of bathrooms)
- acre_lot (Property/Land size in acres)
- city (city name)
- state (state name)
- house_size (house area/size/living space in square feet)
- prev_sold _date (previously sold date)
- price (housing price, it is either the current listing price or recently sold price if the house is sold recently)

Data Cleaning and Preparation

Ensuring the quality and suitability of a dataset for modeling is of paramount importance in the field of data science. These steps encompass thorough data exploration and preprocessing. Specifically, we have undertaken procedures including initial setup and data overview, handling missing values, and data capping with Truncation as part of the data cleaning process.

The data cleaning process began with the initial setup, which involved the integration of essential libraries such as 'pandas,' 'numpy,' 'matplotlib,' 'seaborn,' and 'sklearn.' 'Pandas' was configured to improve data visibility by displaying up to 50 rows and columns, a crucial modification for effective data inspection. The 'head()' function was used to obtain a preliminary view of the data, while 'describe()' was employed to extract detailed statistical insights, facilitating the identification of data distribution characteristics and potential anomalies.

In the "Handling Missing Values" phase, our primary objective was to enhance the dataset's integrity and reliability through a series of meticulous data refinement and imputation steps. Initially, we identified and removed 71 records with empty 'price' values to ensure the accuracy of our predictive model. Subsequently, we filtered the dataset to retain only those records with a 'prev_sold_date,' ensuring that the analyzed prices represented actual transaction values. To address outliers and stabilize input distributions, we capped and floored data at specific percentiles. Various methods, including mean, median filling, and mean encoding, were attempted to handle missing values, with consideration for potential overfitting with mean encoding.

For comprehensive missing value handling, we employed indicator variables such as 'bed_missing' and 'bath_missing' to signify missing data instances, which could be used as interaction terms during modeling. For variables with less than 5% missing data ('bed' and 'bath'), we explored imputation strategies involving mean, median, and even the possible exclusion of such records. 'acre_lot' and 'house_size' missing values were imputed using their respective means, while 'bed' and 'bath' values were replaced with medians. Although we considered category-based imputation, we ultimately discarded this approach due to concerns about oversimplifying the data and introducing bias in representing property diversity within price categories.

To better account for the majority of housing features and prices while mitigating the impact of exceptional properties, we introduced upper and lower limits for specific attributes. Our approach involved setting the upper limit for 'house_size_filled' at the 99th percentile and the lower limit at the 1st percentile, effectively constraining the range between 500 and 7000. Additionally, 'bed_filled,' 'bath_filled,' and 'acre_lot_filled' were also subjected to capping at their respective 99th percentiles. This strategic adjustment aimed to ensure that our predictive model accurately reflected the dynamics of the mainstream housing market, avoiding undue influence from extreme outliers or unique properties that did not align with our primary target audience. By implementing these upper and lower bounds, we improved our ability to capture the typical data distribution, ultimately enhancing the model's stability and reliability.

Furthermore, our analysis of the summary statistics revealed that the response variable 'price' exhibited a wide distribution, ranging from a minimum of 0 to a maximum of 8.75*e+08. Such a broad range presented challenges in building a robust predictive model, especially given the limited input features available. Our objective was to create a model that primarily addressed typical housing transactions, rather than focusing on the extreme ends of the market. To achieve this, we considered either floor and cap settings at the 1% and 99% percentiles for the 'price' variable or the removal of outliers beyond these thresholds. Additionally, during the modeling phase, we planned to explore the option of log-transforming the response variable 'price' to potentially enhance the model's performance.The dataset has been

meticulously optimized and is now well-prepared for subsequent modeling and analysis, ensuring data representativeness and model reliability.

Data Exploratory Data Analysis (EDA)

Using the refined dataset, we conducted an initial EDA on the independent variables, aiming to calculate the average price for each level of these variables.

In our preliminary EDA, we observed an intriguing phenomenon: a significant drop in house prices in February. To investigate whether this trend was prevalent across all states, we segmented the data and found that only New York, Massachusetts, and New Hampshire exhibited this unique pricing pattern.We will conduct further discussions and studies on this observation next.

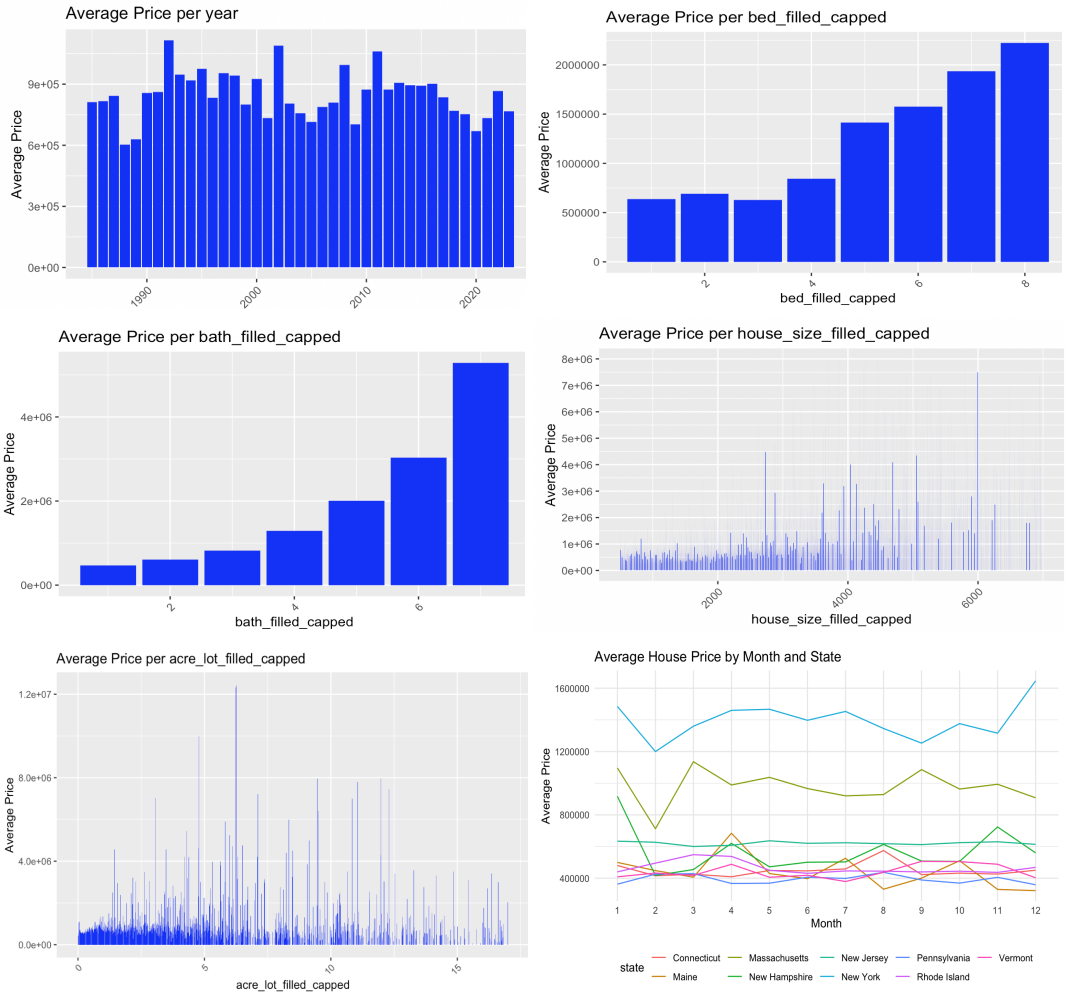


Figure 1: Real Estate Market Data Analysis Overview

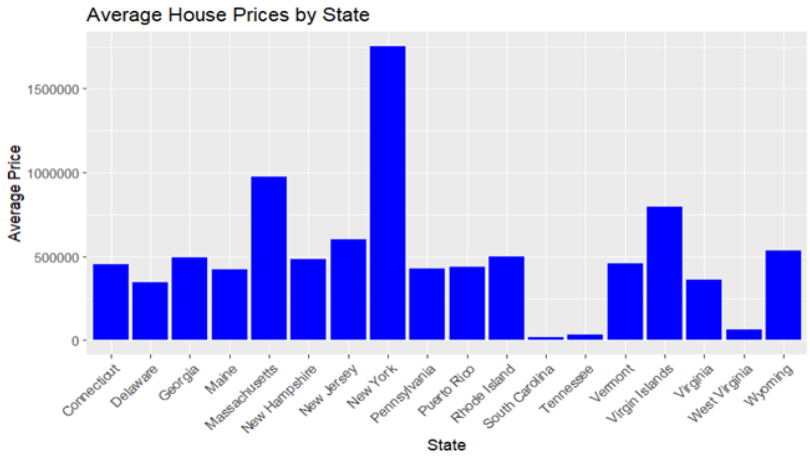


Figure 2: Average House Price by State

To explore the relationship between the predicting variable and the response (price), we utilized the heatmap to visualize how the predicting variable correlated to the response (price), as well as the correlation between each pair of predicting variables. As our goal was to create a price predicting model with the existing predicting factors in the dataset, we first ran an initial multilinear regression with all the original form of variables to test whether our existing data would be able to fulfill the underlying assumption for multilinear regression and whether any transformation of variables would be needed.

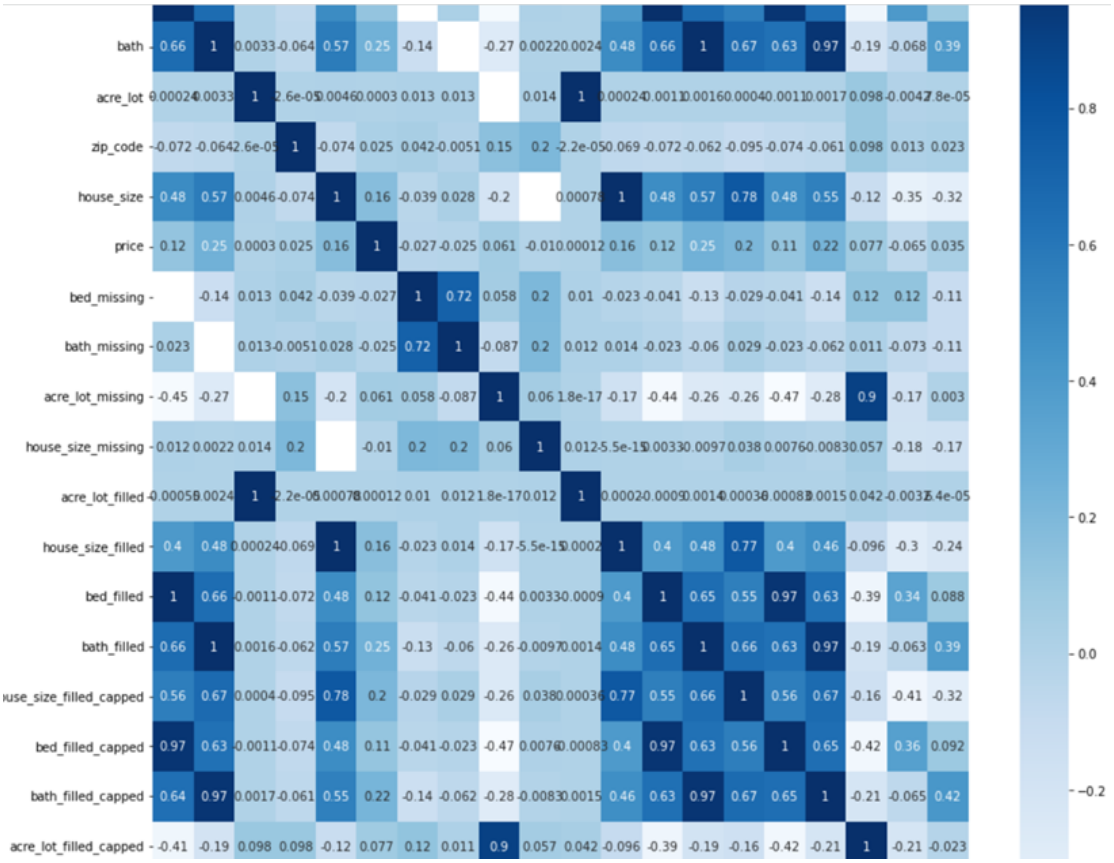
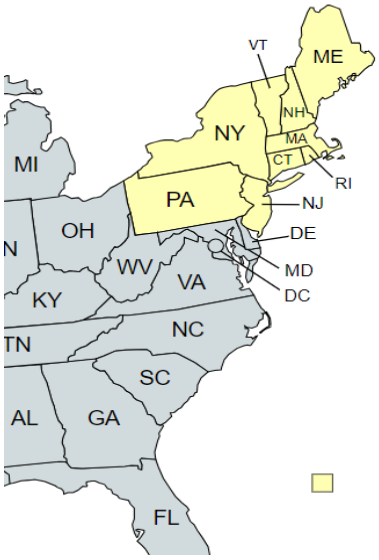


Figure 3: Heatmap Correlation Analysis for Housing Price Predictors

The assumptions for multilinear regression are linearity, constant variance, uncorrelated errors and normality. We plot each predicting variable to the response variable to check for the linear relationship. The variance inflation factor (VIF) from the model also indicated multicollinearity did not exist, which confirmed what we can see from the heatmap. Then we also plot the residual from the initial model to the fitted value to check whether it meets the constant variance criteria. We also used qq-plot and histogram to check for normality.

Since our dataset only contains a handful of features available to use, we tried to use all features at the initial modeling stage. Once each model was built, we then assess if any specific features are not significant to our prediction model respectively.

However, from most of the plots that we got, we noticed an issue regarding price differences among states that are quite significant. As we can see from the bar chart, the average house price of a few states (New York and Massachusetts) in our dataset are at very high end, around $1000k range, compared to some of them are on the lower side of $250k (South Carolina, Tennessee and West Virginia). In order to enhance the robustness and reliability of our model, we opted to omit data from states with limited datasets, as this could lead to increased sensitivity and potentially less accurate results. We focused our analysis on data from Massachusetts, New York, New Jersey, Pennsylvania, Rhode Island, and Vermont. The decision to include these specific states was driven by their more comprehensive data availability, which we anticipate will contribute to the development of more precise and broadly applicable predictive models.



We are discussing the necessity of building models to fit in different price ranges so it can better reflect the income level or purchasing power across different states. And another optional solution is to use State as an categorical input variable to account for the inherent difference between the price of different states.

Feature Engineering

The dataset under consideration encompasses a limited selection of features. Consequently, to enhance the richness of the dataset and facilitate a more comprehensive analysis, we envisage incorporating several additional variables derived from the existing features. In our detailed analysis of the dataset, we are particularly introducing refined features such as 'bed/house_size_filled_capped' and 'bath/house_size_filled_capped'. These features consider the ratio of house size to the number of bedrooms and bathrooms, incorporating data filling and capping to address extreme values or anomalies. Moreover, the 'house_size/acre_lot_filled_capped' feature extends the basic 'house_size/acre_lot' ratio by also including data imputation and capping, providing a more accurate analysis of land use efficiency. Additionally, we have introduced features like 'bed_vs_bath', 'bed/acre_lot_filled_capped', and 'bath/acre_lot_filled_capped', exploring the proportional relationship between the number of bedrooms and bathrooms. This further refines our understanding and analysis of property value.In parallel, we have

also investigated other features, such as sold year and month, to discern the relationship between temporal factors and house price.

In the course of our model training procedure, it is evident that a considerable number of newly incorporated features exhibit statistical significance in explaining the response variable, which is the sale price. Furthermore, during the variable selection phase, these identified variables are deliberately retained in the model based on the criteria established by the model search indicator.

Overall of Modeling

Multilinear Regression

There are a few models commonly used in property pricing. One of them is the Hedonic pricing model (Malpezzi, S. 2003 Hedonic Pricing Models: A Selective and Applied Review). The hedonic pricing model implements the hedonic regression to analyze the supply and demand of a composite good. By estimating how each existing factor influences the price of a property, Hedonic pricing models are relatively straightforward since they use actual market prices and extensive, available data sources. Quantile regression model is another popular choice, Zietz and Sirmans (2007, Determinants of House Prices: A Quantile Regression Approach) pointed out that the homebuyer of higher-priced homes value certain housing characteristics differently from buyers of lower-priced homes, leading to the variation in pricing standard for properties in different price range. However, as our objective is to focus on the factors we obtained from the existing dataset and estimate the extent to which these predictors determine the house price of different locations in the Northeastern United States, we would first select the hedonic pricing model and apply multilinear regression to our data.

We then performed validity analysis to ensure all relevant predictors (independent variables) are included in the model. We use metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to quantify the prediction accuracy and examine the relationships between predictors.

For reliability analysis, we checked for consistency and ensured that the model's reliability is not heavily dependent on a specific analyst's choices. We also perform cross-validation to determine whether the model's performance is stable across different subsets of the dataset.

To conduct a multilinear regression, we begin by loading and exploring data from a Parquet file, focusing on key variables, and constructing an initial multilinear regression model (Model 1). We then delve into examining the relationship between house prices and the "month" variable, conducting residual analysis to pinpoint potential model enhancements. Implementing a logarithmic transformation of the response variable (Model 2), we further improve the model by introducing interaction terms created in feature engineering, enhancing goodness of fit and addressing multicollinearity. Exploiting different model selection criteria by comparing AIC, BIC, adjusted R-squared and F-test result (anova-test) of logarithmic transformed models with varied sets of variables. We then perform Forward-Backward Stepwise regression based on BIC value as a process of model search. BIC tends to penalize complex models more heavily than AIC, it provides a more conservative result. As we notice slightly multicollinearity among a few of predictor variables, we apply Ridge regression on the model to stabilize and improve the model performance. Simultaneously we take advantage of using Lasso regression to make variable selection on the same model. The Elastic Net regression is applied, and the final model's performance is assessed on the validation set. The result of Lasso regression as well as Elastic Net regression do not reduce the amount of variables from Model 2. Additionally, the validation data set has been applied to Ridge

regression, Lasso regression as well as Elastic Net regression to validate the model . We aim to determine the ultimate model and utilize it to predict and evaluate prices on test data, culminating in a Root Mean Squared Error calculation.

According to the result from the cross validation, Model2, which is the same model as what we obtain from the Forward-Backward Stepwise regression performs the best meeting all the screening criteria. We implement this model to make a price prediction with our test data and compare the result to the original listing price by calculating the RMSE. To affirm our multilinear model selection decision , we apply the same testing procedure and make a prediction using the Ridge regression mode, Lasso regression model as well as the Elastic Net regression model. The RMSE of the latter three models perform slightly worse than the Model2. Therefore, Model2 would be our most appropriate multilinear model in terms of goodness of fit and prediction power.

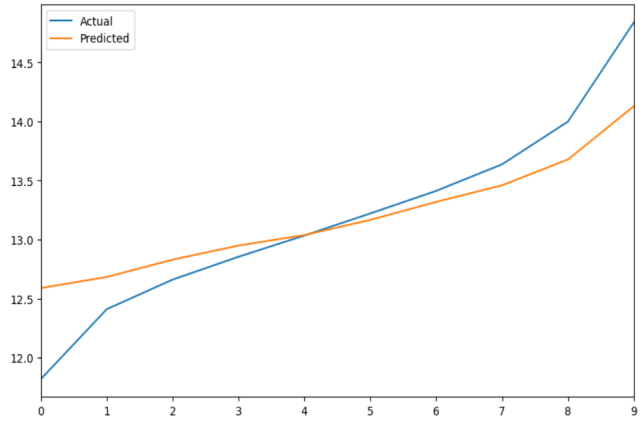
```
                                  5.662e+00  2.521e-01  22.456  < 2e-16 ***
stateMaine                        6.813e-03  1.063e-02   0.641   0.5218
stateMassachusetts                6.850e-01  4.279e-03 160.096  < 2e-16 ***
stateNew Hampshire                2.273e-01  7.204e-03  31.556  < 2e-16 ***
stateNew Jersey                   3.419e-01  3.715e-03  92.052  < 2e-16 ***
stateNew York                     1.112e+00  4.130e-03 269.157  < 2e-16 ***
statePennsylvania                 8.797e-02  7.881e-03  11.163  < 2e-16 ***
stateRhode Island                 2.979e-01  6.851e-03  43.484  < 2e-16 ***
stateVermont                     -1.282e-01  7.841e-03 -16.349  < 2e-16 ***
year                              3.067e-03  1.253e-04  24.478  < 2e-16 ***
month2                           -2.470e-02  5.974e-03  -4.135 3.55e-05 ***
month3                           -1.121e-03  5.736e-03  -0.195   0.8450
month4                            5.903e-03  5.712e-03   1.033   0.3014
month5                            2.377e-02  5.607e-03   4.239 2.24e-05 ***
month6                            9.551e-03  5.421e-03   1.762   0.0781 .
month7                            2.640e-02  5.385e-03   4.901 9.52e-07 ***
month8                            3.535e-04  5.394e-03   0.066   0.9477
month9                            3.924e-03  5.539e-03   0.708   0.4787
month10                          -9.025e-03  5.496e-03  -1.642   0.1005
month11                           1.452e-02  5.666e-03   2.563   0.0104 *
month12                          -3.477e-03  5.551e-03  -0.626   0.5311
house_size_filled_capped          1.478e-05  3.122e-06   4.735 2.20e-06 ***
bed_filled_capped                -2.502e-02  2.316e-03 -10.805  < 2e-16 ***
bath_filled_capped                4.439e-01  3.536e-03 125.528  < 2e-16 ***
acre_lot_filled_capped            1.760e-02  2.973e-04  59.198  < 2e-16 ***
`bed/house_size_filled_capped`    1.716e+01  3.042e+00   5.641 1.69e-08 ***
`bath/house_size_filled_capped`  -1.903e+02  6.153e+00 -30.930  < 2e-16 ***
bed_vs_bathbed_less_than_bath     2.460e-02  5.104e-03   4.820 1.43e-06 ***
bed_vs_bathbed_more_than_bath    -2.912e-02  3.800e-03  -7.663 1.82e-14 ***
`bed/acre_lot_filled_capped`      3.958e-04  9.300e-05   4.256 2.08e-05 ***
`bath/acre_lot_filled_capped`    -3.281e-03  1.326e-04 -24.745  < 2e-16 ***
`house_size/acre_lot_filled_capped`  2.862e-06  1.584e-07  18.066  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.566 on 262857 degrees of freedom
Multiple R-squared:  0.5438,    Adjusted R-squared:  0.5437
F-statistic: 1.011e+04 on 31 and 262857 DF,  p-value: < 2.2e-16
```

Table 1: Model 2 summary

Table II: Prediction Performance

models	RMSE
Full model(Model 2)	0.5678201
Stepwise	0.5678201
Elastic net model	0.6734284
Lasso model	0.6734353
Ridge model	0.6757534



	bin	actual_col	pred_col	% diff
0	(9.942, 12.252]	11.820190	12.590304	6.515248
1	(12.252, 12.542]	12.411578	12.684252	2.196932
2	(12.542, 12.765]	12.661933	12.830596	1.332051
3	(12.765, 12.946]	12.855143	12.949779	0.736168
4	(12.946, 13.12]	13.034845	13.038915	0.031227
5	(13.12, 13.305]	13.221583	13.167272	-0.410771
6	(13.305, 13.521]	13.412605	13.319431	-0.694680
7	(13.521, 13.764]	13.637333	13.460011	-1.300269
8	(13.764, 14.221]	13.998562	13.678168	-2.288762
9	(14.221, 20.59]	14.838075	14.128802	-4.780088

Figure 4: Actual price vs. Predicted price of the test dataset on Random Forest model

Random Forest Regression

In addition to the multilinear regression model, we explored the application of a random forest regression (RF) for predicting house prices. RF is a robust and versatile algorithm that is particularly suitable for predicting house prices due to its ability to handle complex, non-linear data relationships and its strong predictive performance. To prepare the dataset for effective modeling, we performed one-hot encoding on categorical features, enabling their integration into the model seamlessly. The Random Forest Regressor was initialized with default parameters, and a fixed random seed was set to ensure reproducibility throughout the training process. Subsequently, the model was trained using the encoded training dataset. To assess its performance, we computed the RMSE by comparing the predicted house prices to the actual target values using a separate validation dataset. Next, we applied the trained model to generate predictions for a previously unseen test dataset, utilizing the same one-hot encoding technique. The RMSE was once again calculated to evaluate the model's performance on this novel data. This comprehensive methodology establishes a baseline for house price prediction and forms a solid foundation for further model refinement and enhancement.

The results from the RF model with default parameters and a random_state set to 123 reveal a clear pattern of overfitting. This is evident in the RMSE values: the RMSE is high on the training data (331,364), slightly higher on the validation set (361,830), and significantly higher on the test set (461,641). The consistent increase in RMSE from the training to the test set indicates that the model is too complex and is fitting noise in the training data rather than capturing the true underlying patterns in the data. In particular, the substantial jump in RMSE on the test set strongly suggests overfitting.

To address this issue, several strategies can be employed. One approach is to reduce the model's complexity, which can be achieved by adjusting parameters like ‘n_estimators’ and ‘max_features’. Currently, with ‘n_estimators’ set at 100 and ‘max_features’ allowing each split to consider all features, there is room to simplify the model for better performance. Additionally, enhancing the feature set through feature engineering could significantly improve results. Combining these steps with the use of cross-validation techniques will provide a more comprehensive understanding of the model's performance across different data subsets. Furthermore, conducting a detailed analysis of feature importance and residual plots would offer deeper insights into the model's behavior and guide more focused improvements.

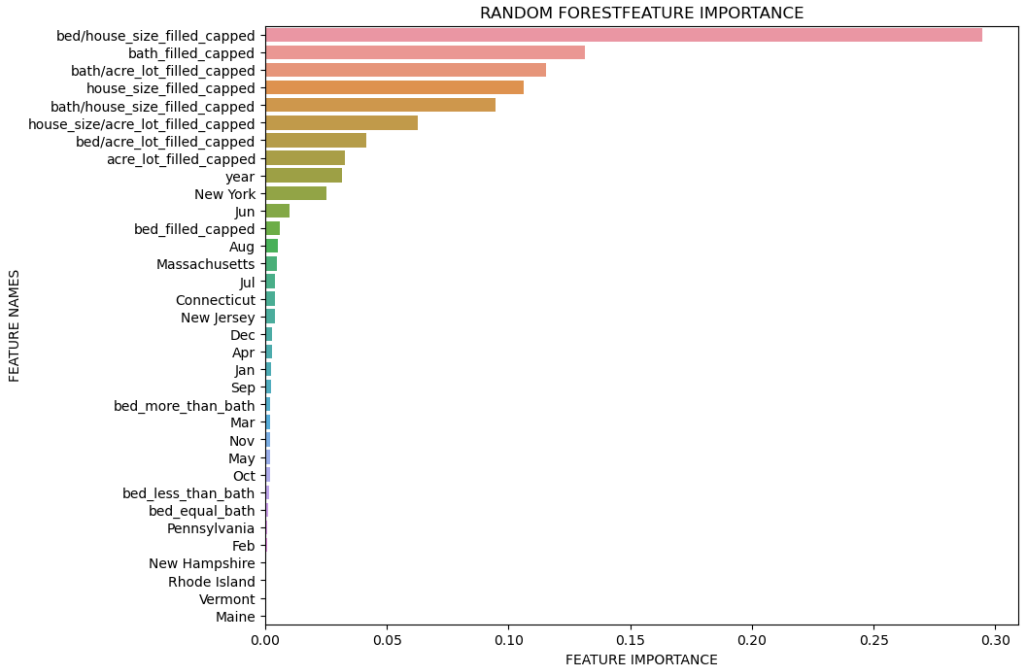
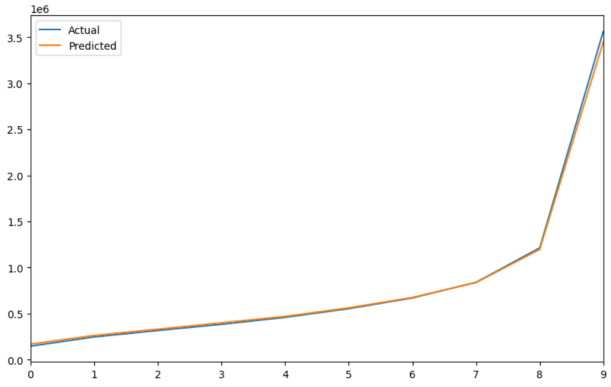


Figure 5: Feature importance ranking in Random Forest Regression



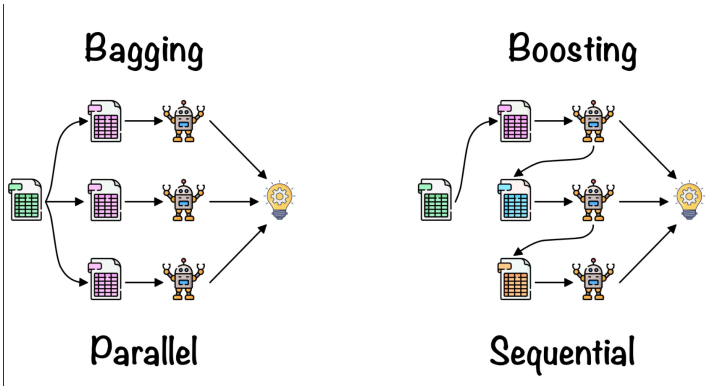
	bin	price	predicted_price	% diff
0	(20799.999, 209500.0]	1.462774e+05	1.687961e+05	15.394545
1	(209500.0, 279900.0]	2.465207e+05	2.626597e+05	6.546713
2	(279900.0, 349900.0]	3.161726e+05	3.306851e+05	4.590058
3	(349900.0, 419000.0]	3.832265e+05	3.999082e+05	4.352947
4	(419000.0, 499000.0]	4.587775e+05	4.693331e+05	2.300811
5	(499000.0, 599900.0]	5.530814e+05	5.621984e+05	1.648393
6	(599900.0, 744900.0]	6.694738e+05	6.746820e+05	0.777946
7	(744900.0, 950000.0]	8.390530e+05	8.367204e+05	-0.278011
8	(950000.0, 1500000.0]	1.212641e+06	1.195899e+06	-1.380658
9	(1500000.0, 875000000.0]	3.564450e+06	3.450772e+06	-3.189193

Figure 6: Actual price vs. Predicted price of the test dataset on Random Forest model

LightGBM family

Lightgbm_v1

In contrast to the Random Forest, which is a bagging algorithm, we also developed a boosting version of a tree-based model using LightGBM.



(https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422)

Unlike Random Forest, LightGBM provides the flexibility to fine-tune more parameters to improve predictive accuracy. However, as we delved deeper into hyperparameter tuning, we encountered overfitting issues. Leveraging Optuna to optimize the hyperparameter tuning process, we successfully achieved a lower RMSE on the validation dataset. Moreover, the performance of the LightGBM model outperformed the Random Forest model when evaluated on the testing dataset as well.

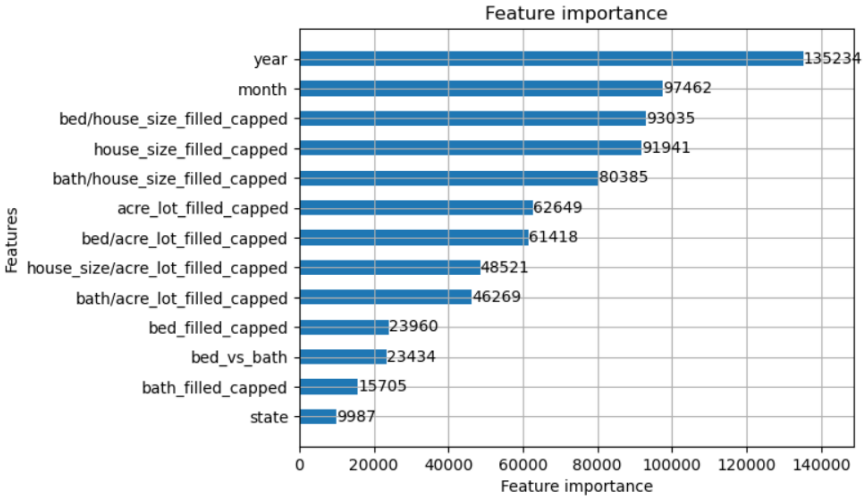
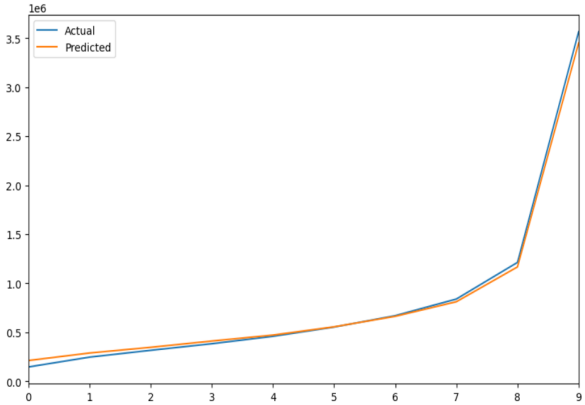


Figure 7: Feature importance ranking in LighGBM_v1



	bin	price	predicted_price	% diff
0	(20799.999, 209500.0]	1.462774e+05	2.119598e+05	44.902620
1	(209500.0, 279900.0]	2.465207e+05	2.886066e+05	17.071943
2	(279900.0, 349900.0]	3.161726e+05	3.467329e+05	9.665710
3	(349900.0, 419000.0]	3.832265e+05	4.107231e+05	7.175004
4	(419000.0, 499000.0]	4.587775e+05	4.716503e+05	2.805905
5	(499000.0, 599900.0]	5.530814e+05	5.557674e+05	0.485631
6	(599900.0, 744900.0]	6.694738e+05	6.625117e+05	-1.039934
7	(744900.0, 950000.0]	8.390530e+05	8.113934e+05	-3.296530
8	(950000.0, 1500000.0]	1.212641e+06	1.166611e+06	-3.795806
9	(1500000.0, 875000000.0]	3.564450e+06	3.449748e+06	-3.217919

Figure 7: Actual price vs. Predicted price of the test dataset on LightGBM_v1

LightGBM_v2

The LightGBM model has consistently provided us with the best predictive performance thus far. In an attempt to further enhance our modeling approach, we decided to build another LightGBM model with a unique twist. This time, we excluded the 'Year' and 'Month' features from our dataset. Our dataset primarily consists of two distinct types of records: those with sold dates and those without. We categorize the records without sold dates as listing records, where the listed price represents the property's price. Our objective was to train the model on the sold records without the 'Year' and 'Month' information, effectively treating them as if we didn't know the precise timing of the sale. Subsequently, we used this trained model to predict the listing prices, allowing us to assess the reasonability of these listing prices.

As anticipated, the removal of the 'Year' and 'Month' variables led to a significant increase in the model's RMSE. However, despite this, we were able to achieve a relatively stable model, with the RMSE remaining fairly consistent across three different data partitions.

Furthermore, we conducted an analysis of the distribution of the differences between the predicted price and the listing price for the listing records, categorized by state. The percentage difference was calculated as (predicted price) / (listing price) - 100%. Here is a summary of our findings:

grouped_percentage_difference	Connecticut	Maine	Massachusetts	New Hampshire	New Jersey	New York	Pennsylvania	Rhode Island	Vermont
(<-100%)	1.3	3.4	2.1	1.7	2.0	10.9	0.8	0.9	2.9
(-100%,-50%)	3.5	9.1	13.0	8.1	7.5	27.7	5.2	5.5	9.4
(-50%, -20%)	8.6	11.8	16.1	13.3	16.6	11.8	14.0	10.6	13.2
(-20%, -10%)	5.6	4.6	5.1	4.9	7.9	4.0	6.7	7.4	4.0
(-10%, 0%)	6.7	4.2	5.3	6.2	6.7	3.8	6.0	10.7	3.4
(0%,10%)	6.5	4.2	5.5	4.2	6.4	2.5	5.9	9.7	3.2
(10%,20%)	8.0	3.3	4.7	3.5	5.2	2.6	4.6	8.2	3.4
(20%,50%)	12.6	8.8	10.6	10.3	12.3	6.5	10.8	13.5	8.4
(50%,100%)	11.8	10.2	10.0	9.8	11.1	6.2	10.0	9.6	9.0
(100%,200%)	10.4	10.9	8.3	9.2	9.2	8.8	9.0	8.8	11.1
(200%+)	25.0	29.5	19.3	28.9	15.0	15.1	26.8	15.3	31.9
Grand Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Grouped range	Connecticut	Maine	Massachusetts	New Hampshire	New Jersey	New York	Pennsylvania	Rhode Island	Vermont
[-20%,+50%]	39.41	25.06	31.19	29.15	38.53	19.47	34.08	49.44	22.46
[-50%,+50%]	48.04	36.87	47.25	42.43	55.08	31.31	48.07	60.09	35.71
[-20%,+20%]	26.80	16.30	20.59	18.80	26.21	12.92	23.25	35.93	14.02
[-50%,+20%]	35.43	28.11	36.64	32.08	42.76	24.77	37.24	46.58	27.26

Table 3: Distribution of Predicted vs. Actual Listing Price Percentage Differences by State

Our focus was on identifying which state exhibited a smaller percentage difference between the predicted sold price and the listing price. This analysis differs from our previous approach, where we examined RMSE by state, which provided an overall distribution of records per state. In this exercise, our goal was to determine if any specific state tended to set more reasonable (close to predicted) listing prices. We explored various grouped ranges, such as -50% to +20%, among others. Notably, the states of 'Connecticut, New Jersey, Pennsylvania, and Rhode Island' exhibited a higher concentration around a 0% difference. This leads us to conclude that using the models' predicted price as the final sale price, the listing prices in these four mentioned states tend to be more reasonable when compared to other states.

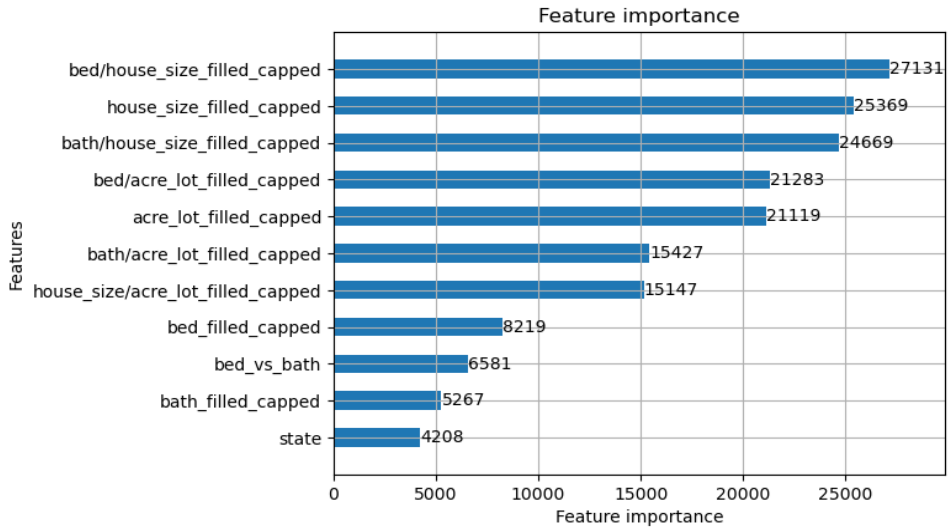
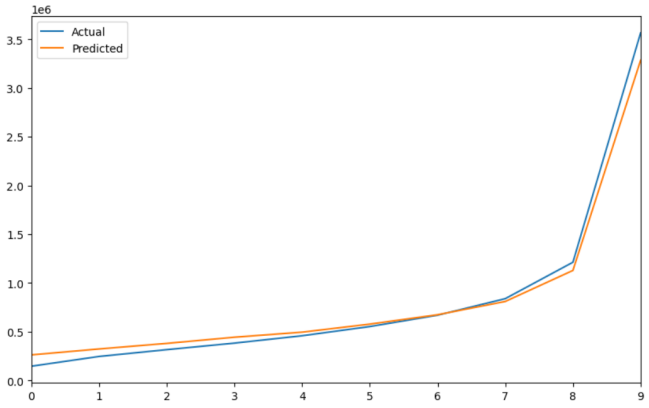


Figure 8: Feature importance ranking in LighGBM_v2



	bin	price	predicted_price	% diff
0	(20799.999, 209500.0]	1.462774e+05	2.625264e+05	79.471665
1	(209500.0, 279900.0]	2.465207e+05	3.233914e+05	31.182256
2	(279900.0, 349900.0]	3.161726e+05	3.804309e+05	20.323811
3	(349900.0, 419000.0]	3.832265e+05	4.438092e+05	15.808586
4	(419000.0, 499000.0]	4.587775e+05	4.955575e+05	8.016969
5	(499000.0, 599900.0]	5.530814e+05	5.782435e+05	4.549426
6	(599900.0, 744900.0]	6.694738e+05	6.742658e+05	0.715792
7	(744900.0, 950000.0]	8.390530e+05	8.101875e+05	-3.440250
8	(950000.0, 1500000.0]	1.212641e+06	1.128525e+06	-6.936584
9	(1500000.0, 875000000.0]	3.564450e+06	3.283928e+06	-7.869981

Figure 8: Actual price vs. Predicted price of the test dataset on LightGBM_v2

Support Vector Regression (SVR)

We also experimented with a SVR model for predicting house prices, selected due to its proficiency in managing intricate, non-linear data correlations.The initial stage of data preprocessing involved the one-hot encoding of categorical variables, a critical step to render them suitable for regression. Subsequently, we applied feature scaling via scikit-learn's 'StandardScaler' to standardize the data, ensuring that each feature contributed equitably to the model's predictions.

The training phase of the model was marked by a thorough tuning of hyperparameters using GridSearchCV. Key parameters like regularization strength ('C'), epsilon, the kernel type, and gamma were meticulously adjusted. To further solidify the model's robustness, a 5-fold cross-validation approach was adopted. We assessed the SVR model's efficacy through a diverse array of metrics, including Mean Squared Error (MSE), RMSE, R-squared (R2), Mean Squared Logarithmic Error (MSLE), MAE, and Mean Absolute Percentage Error (MAPE).

The SVR model's performance indicated significant challenges in accurately predicting house prices. High MSE and RMSE values on both the training and validation sets suggested substantial deviations of the model's predictions from the actual values. Particularly concerning were the negative R2 scores, with -0.008 on the training set and -0.027 on the validation set, indicating that the model performed worse than a simple baseline that would always predict the mean value of the target variable.

Additionally, the high values in MSLE, MAE, and MAPE metrics further confirmed the model's limited predictive accuracy. The best parameters identified from the GridSearchCV were {'C': 0.1, 'epsilon': 0.01, 'gamma': 'scale', 'kernel': 'linear'}, pointing towards a simpler model that might have contributed to underfitting.

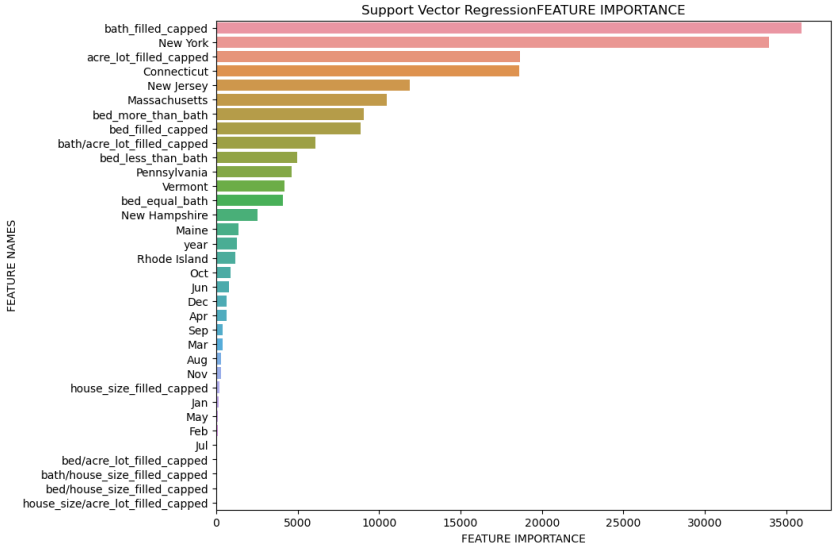
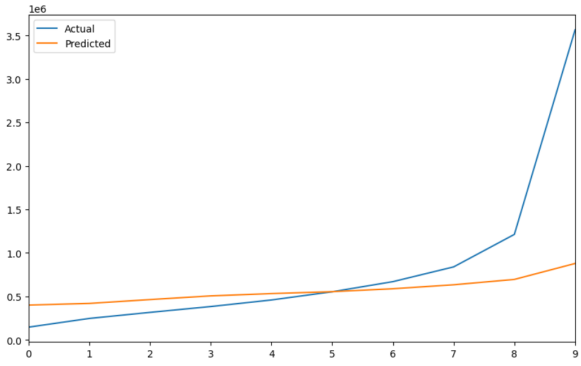


Figure 9: Feature importance ranking in SVR



	bin	price	predicted_price	% diff
0	(20799.999, 209500.0]	1.462774e+05	400091.584408	173.515689
1	(209500.0, 279900.0]	2.465207e+05	419230.021567	70.058735
2	(279900.0, 349900.0]	3.161726e+05	463381.734159	46.559744
3	(349900.0, 419000.0]	3.832265e+05	505725.128038	31.965057
4	(419000.0, 499000.0]	4.587775e+05	532236.231620	16.011848
5	(499000.0, 599900.0]	5.530814e+05	554051.960383	0.175477
6	(599900.0, 744900.0]	6.694738e+05	587817.833913	-12.197038
7	(744900.0, 950000.0]	8.390530e+05	633543.214441	-24.493065
8	(950000.0, 1500000.0]	1.212641e+06	694930.034536	-42.692845
9	(1500000.0, 875000000.0]	3.564450e+06	878378.114519	-75.357258

Figure 10: Actual price vs. Predicted price of the test dataset on SVR

Model Performance Comparison

RMSE is favored for model comparison as it directly reflects the average error in the predictions. This metric is particularly effective in the context of real estate, where precise value estimates are vital, and it

ensures that larger errors are given more weight, underscoring the importance of accuracy in high-stakes predictions.

Root Mean Square Error comparisons for each model are summarized as follows:

Model	RMSE		
	Training	Validation	Test
Multiple Linear Regression - Log Scale	2,695,397	1,248,450	3,173,411
SVR	2,804,160	1,468,526	3,259,656
Random forest	331,364	361,830	461,641
LightGBM	114,888	353,323	324,648
LightGBM - without Year/Month - Can be used to predict listing price	512,039	587,795	548,878

The comparison of models using RMSE reveals that LightGBM outperforms others with the lowest errors, indicating superior prediction accuracy and generalizability for house pricing. Multiple Linear Regression and SVR show moderate to high errors, respectively, with SVR being the least accurate. Random Forest, although low on training error, suggests underfitting with higher test RMSE. LightGBM without Year/Month data increases RMSE across 3 data partitions but still remains a strong model, positioning standard LightGBM as the most suitable for precise real estate price predictions.

Through comparing actual and predicted house prices across different price ranges in training, validation, and test sets, it was found that although there are discrepancies in accuracy among the models, they all tend to perform with higher relative accuracy within the price range of $400,000 to $1,000,000. In the assessment of predictive models, SVR showed the greatest difference between predicted and actual prices, suggesting lower precision. The RF model demonstrated the least difference on test data but was prone to overfitting, as indicated by a significant increase in error compared to the training data. The LightGBM model, while less accurate for lower-priced houses, displayed consistent accuracy across other price segments and maintained uniform performance across all datasets, highlighting its overall stability.

Overall, LightGBM was chosen as the preferred house price prediction model due to its consistently lower RMSE errors across various datasets, demonstrating superior predictive accuracy. Additionally, it displayed remarkable stability and relatively high accuracy within the crucial price range of $400,000 to $1,000,000, which is highly relevant in the real estate market. Even when excluding certain data features, LightGBM remained a robust and dependable choice for precise real estate price predictions, making it the optimal model for this study.

The model can be serialized (pickled) to meet the specific requirements of the target audience. With feature engineering integrated into the pipeline, the model can be swiftly transitioned into production.

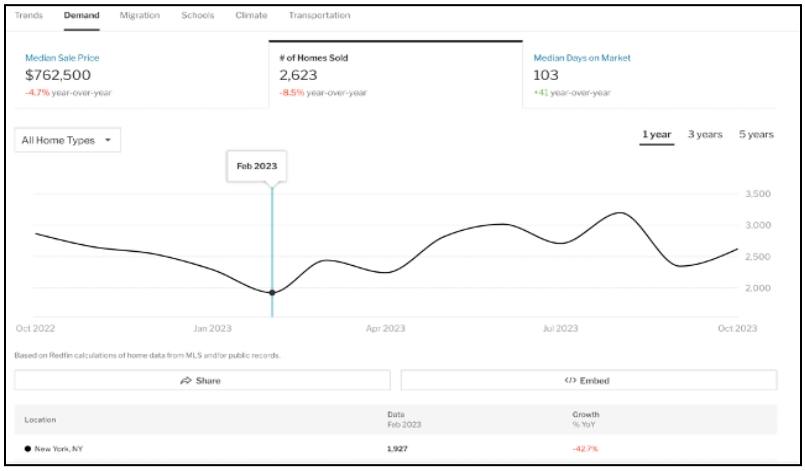
Discussion

The multilinear regression model, though less accurate in predicting home prices compared to more complex models like random forest or gradient boosting, offers a distinct advantage in interpretability. While the LightGBM model excels at capturing intricate data patterns for precise predictions, the multilinear model allows us to unravel the nuanced relationships inherent in the data. It represents a deliberate tradeoff between the clarity of interpretation and the pursuit of exact prediction. In practical

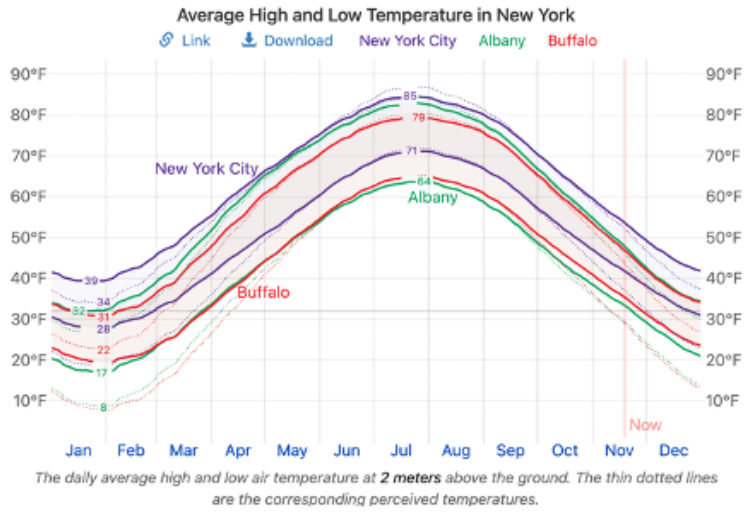
terms, the LightGBM model would serve as the primary tool here for making home price predictions due to its capacity to comprehend complex patterns. However, the multilinear model remains indispensable for shedding light on the underlying relationships among certain parameters used in our modeling. It not only provides valuable insights for interpretability but also offers strategic decision-making advice, enhancing our understanding of the factors influencing home prices.

Upon examining the coefficients derived from the multilinear model, a few facts from the data should be noted. The predictor "year" as a quantitative predictor, stands out as a conditional inflation-adjusted factor, as evidenced by its coefficient of 0.003067. This implies that, on average, adding one year to the timeline corresponds to a 0.3067% increase in house prices, while holding all other variables constant. It is noted the minimum year in our dataset is 1985.

Turning attention to the categorical variable "month", we observed a significant decline in real estate sale prices for the month of February in several states, including New York, Massachusetts, and New Hampshire, while other states did not experience a similar trend. The same pattern was observed from analyzing the data from Redfin (https://www.redfin.com), confirming this observation.



This pattern is intriguing and may suggest that February could be an opportune time to buy a home in these states. Several factors might contribute to this price drop. Firstly, February is the coldest month in these three states, as indicated by weather data in New York (https://weatherspark.com/countries/US/NY). The harsh weather conditions may discourage people from venturing out to view properties, and the winter landscape may not showcase homes as attractively. Secondly, the holiday season, including Christmas and New Year, precedes February, and individuals may have allocated significant funds during this period, affecting their home-buying capacity. Thirdly, families often prefer to move at the end of the school year to avoid disrupting their children's education, making late spring and early summer more popular for home purchases than winter. As suggested by prior research, the price decline is likely influenced by a shift in demand for space. Other contributing factors may include tax planning and changes in marketing activities, as sellers might be more inclined to list their properties in the spring when homes look more appealing, and there is increased buyer activity. While it's challenging to definitively conclude the reason for the February price drop, this intriguing trend could provide valuable insights for both buyers and sellers.



Given our dataset's focus on home sale prices from several states along the east coast of the United States, Connecticut serves as the baseline for the "State" parameter. The multilinear regression model facilitates a comparative analysis of price differences across these states. With the exception of Vermont, all other states demonstrate home prices surpassing those in Connecticut. Notably, New York stands out, demonstrating a remarkable price difference, with home prices exceeding those in Connecticut by over 100% despite similar characteristics.

In essence, the multilinear regression model not only quantifies the impact of temporal and categorical variables on home prices but also unveils intriguing regional dynamics, shedding light on the considerable variation in housing markets across the Northeast of the United States.

Therefore, employing various models for analysis is crucial, as no single model is flawless. The selection of a model should align with the specific objectives of the research and be meticulously tailored to those needs. In our study, the LightGBM model yielded high predictive accuracy, while the multilinear model afforded deeper insights into the dataset. Moving forward, we will exercise careful consideration in model selection to minimize bias and maximize the potential for uncovering significant insights.

Conclusion and Impact

Our recommendation after analysis is to employ a dual-model strategy. First, we advocate for the utilization of a Multilinear Regression model, incorporating all available parameters from the initial dataset. Its purpose is to illuminate the distinctions in home prices across various locations in the northeastern United States. This model serves as a valuable tool for gaining insights into regional variations.

Conversely, the LightGBM model family assumes the role of our primary predictive tool for accurate sale price estimations. This model family, as delineated earlier, comprises both the LightGBM model incorporating the "month" and "year" parameters and its counterpart without these temporal variables. The training and validation of these models were conducted using historical sales data, subsequently tested against listing price data. The predictive outcomes of this model family hold utility in furnishing sale price estimates for prospective houses anticipated to enter the market shortly. Furthermore, these predictions extend to encompass estimated prices for both properties with historical sales records and those without such documented transactional history.

This dual-model strategy seeks to balance interpretability and precision, leveraging the explanatory power of the Multilinear Regression model alongside the predictive accuracy of the LightGBM model family. It caters to both immediate pricing needs for upcoming listings and broader forecasting requirements for diverse real estate scenarios.

Our model holds significant potential for both public and private sectors. Within the governmental sphere, the application of our multilinear regression model could facilitate an in-depth examination of home price disparities in the northeastern region of the United States. Subsequently, local authorities can leverage the insights garnered to implement targeted fiscal policies, such as adjustments in property taxes or related sales taxes, to strategically stimulate or regulate the local real estate market.

Moreover, the model serves as a valuable instrument for monitoring real estate market dynamics, functioning as an effective gauge of both supply and demand. Additionally, it can serve as a reliable indicator of general inflation trends in home prices and construction material costs. Institutions contemplating new operations in the northeastern U.S. can benefit from the model's capacity to offer comparative analyses of housing costs across different locations. This information proves essential for assessing living expenses and resource availability in prospective regions, particularly considering the impact of housing costs on employee commuting times and associated wage considerations.

Real estate companies can capitalize on our model by integrating it into their platforms to develop predictive tools. For instance, they may create a home price prediction application or calculator on their websites. This tool empowers potential homebuyers to estimate property prices in various states or make detailed comparisons based on specific property characteristics. This strategic integration enables companies to enhance user experience and provide valuable insights for decision-making.

Individual home buyers and investors can also derive substantial benefits from our models. When contemplating the purchase of a home or considering investment opportunities, they can leverage our insights to gain a deeper understanding of the key factors influencing home prices. For instance, as discussed earlier, the observation of a price drop in February may offer valuable insights for home buyers and sellers. They can also employ our model as a resource for estimating offer prices during real estate transactions. Furthermore, it functions as a vigilant monitoring tool for identifying property investment opportunities. Armed with this knowledge, individuals can make more informed budgetary decisions, significantly enhancing their ability to make sound and profitable investments. In essence, our model serves as a versatile and comprehensive resource with far-reaching implications for both governmental policymaking and private sector decision support within the real estate domain.

Further Research

Several limitations are inherent in our model, warranting careful consideration. Firstly, the model operates within the constraint of not accounting for unanticipated perturbations in the economic or political landscape. Notably, the dynamic nature of home prices is susceptible to influences stemming from fiscal policy alterations, such as changes in property taxes, and broader economic factors encompassing interest rates, employment levels, and overall economic growth. These external factors, being pivotal determinants in shaping home prices, pose a challenge to the model's assumptions, as it is predicated on the absence of abrupt shifts in the prevailing economic or fiscal environment.

Secondly, the disparity in home prices across states, as captured by our model, serves primarily as a general indicator for comparative analyses. Notably, the model's architecture lacks the granularity inherent in incorporating specific zip codes or area zoning variables. Consequently, the home price evaluations generated by the model are approximations representing average prices for houses

characterized by specific features. This assessment does not account for the nuanced influence of specific zoning regulations or area-specific factors that may exert differential impacts on property values within distinct geographical locales. The model, designed to provide a broad perspective on home price differentials among states, operates on the premise of characterizing average pricing based on inherent property characteristics, without delving into the localized intricacies introduced by zoning regulations or other area-specific considerations.

Moreover, the model exhibits constraints in predicting sale prices for properties characterized by extreme values, exemplified by vacant land, properties under construction, or luxury residences. The model is trained on sale price within a specific range, specifically from $20,800 to $875,000,000. Consequently, any deviation beyond this delineated range falls outside the model's scope, leading to extrapolation challenges.

Lastly, the temporal scope of our model is oriented towards short-term predictions, encompassing the near-term horizon of 1 to 5 years or retrospective estimates for home prices listed within the past 1 to 5 years. Recognizing the ever-changing factors influencing home prices, it is advisable to exercise caution and periodic reevaluation. Continuous model retraining, coupled with adjustments informed by new data, is recommended to ensure the sustained accuracy of estimation and prediction outcomes. This iterative approach acknowledges the inherent volatility in the real estate landscape and serves to enhance the model's adaptability to evolving market dynamics.

Given the constrained computational resources and a limited set of features at our disposal, our model inherently harbors certain limitations. The augmentation of available features has the potential to substantially enhance the predictive capacity of the model.

For instance, an expansion of features to include precise property addresses could enable the derivation of geographic coordinates, thereby providing access to pertinent geographical and demographic information. Factors such as proximity to educational institutions, distance to major thoroughfares, and census data reflecting neighborhood characteristics could contribute valuable insights. Incorporating essential property-related attributes, such as the original construction year, renovation details, and the number of parking spaces (both detached and attached), would further enrich the model's predictive capabilities.

The prospect of integrating data from platforms such as Zillow for model development represents an avenue for leveraging additional information. With heightened computational resources, comprehensive datasets could be harnessed to refine the predictive accuracy of the model significantly.

Furthermore, the incorporation of recent sales records from the neighborhood, processed and linked to individual property listings, holds potential for providing homebuyers with a more comprehensive understanding of recent transactions within the vicinity. This strategic augmentation aligns with common homebuyer practices, as individuals often reference recent sales records in the same neighborhood to inform their property valuation assessments.

In summary, an expansion of features, coupled with increased computational resources, has the capacity to propel the model's predictive efficacy, offering a more nuanced and comprehensive representation of real estate dynamics.

References

Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review, in Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan, T.O. Sullivan and K. Gibbs (Eds.), Blackwell.

Epple, D. (1987). Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products, Journal of Political Economy, 95, 59-80.

Zietz,J. Zietz,E and Sirmans, G.(2008),Determinants of House Prices: A Quantile Regression Approach, The Journal of Real Estate Finance and Economics

Antonin Bergeaud, Jean-Benoît Eyméoud, Thomas Garcia, Dorian Henricot. Working from home and corporate real estate. Regional Science and Urban Economics, Volume 99, 2023, 103878, ISSN 0166-0462

Milcheva, S. Volatility and the Cross-Section of Real Estate Equity Returns during Covid-19. J Real Estate Finan Econ 65, 293–320 (2022)

Gale, H., Roy, S.S. Optimization of United States Residential Real Estate Investment through Geospatial Analysis and Market Timing. Appl. Spatial Analysis 16, 315–328 (2023).

Guangjie Liu, "Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model", Scientific Programming, vol. 2022, Article ID 5750354, 8 pages, 2022.

Appendix

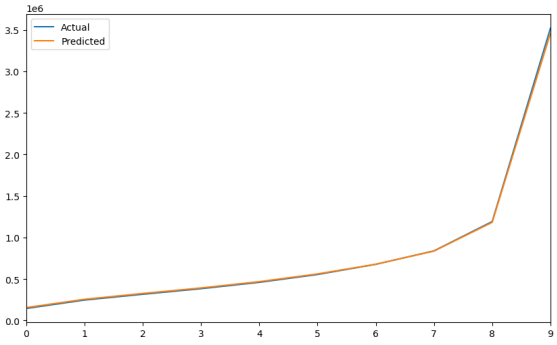
Table S1: Predictors and descriptions/units in cleaned data set

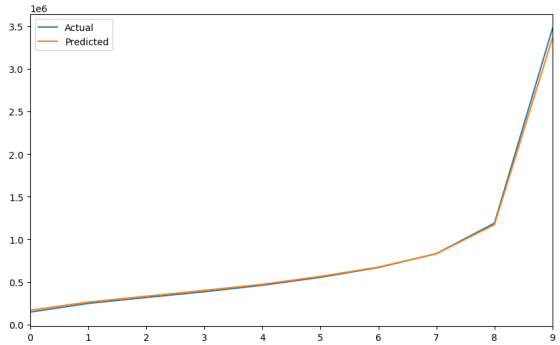
Variable	**Description/Units**	**Variable**	**Description/Units**
price	Sold price	bed/house_size_filled_capped	Ration of bedrooms numbers to the living area square ft
state	House location	bath/house_size_filled_capped	Ration of bathrooms numbers to the living area square ft

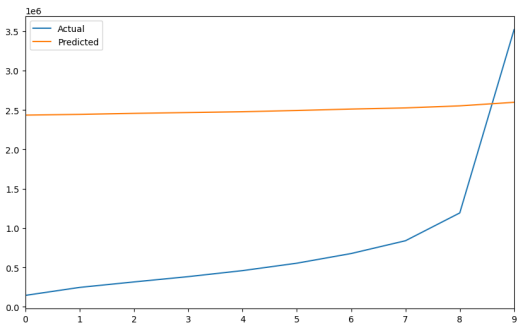
bed_vs_bath	Ratio of the number of bedrooms to bathrooms.	year	Year of sale
house_size_filled_capped	The square ft of each house after missing data filled and removed top and end 1% of house size	month	Month of sale
bed_filled_capped	Total number of bedrooms in the house, after missing data filled and removed outliers	bed/acre_lot_filled_capped	Ratio of bedrooms numbers to the size of the house lot in acre
bath_filled_capped	Total number of bathrooms in the house, after missing data filled and removed outliers	bath/acre_lot_filled_capped	Ratio of bathrooms numbers to the size of the house lot in acre
acre_lot_filled_capped	Size of the house lot in acre, after missing data filled and removed outliers	house_size/acre_lot_filled_capped	Ratio of the living area square footage to the lot size in acres

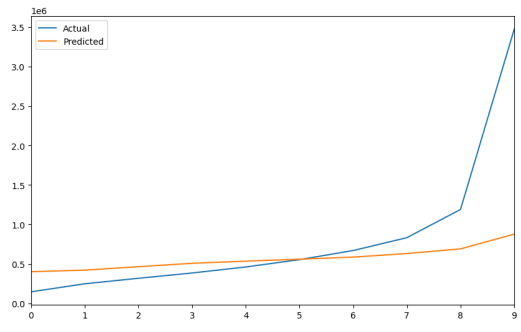
	status	bed	bath	acre_lot	city	state	zip_code	house_size	prev_sold_date	price
0	for_sale	3.0	2.0	0.12	Adjuntas	Puerto Rico	601.0	920.0	NaN	105000.0
1	for_sale	4.0	2.0	0.08	Adjuntas	Puerto Rico	601.0	1527.0	NaN	80000.0
2	for_sale	2.0	1.0	0.15	Juana Diaz	Puerto Rico	795.0	748.0	NaN	67000.0
3	for_sale	4.0	2.0	0.10	Ponce	Puerto Rico	731.0	1800.0	NaN	145000.0
4	for_sale	6.0	2.0	0.05	Mayaguez	Puerto Rico	680.0	NaN	NaN	65000.0

Figure S1: Actual price vs. Predicted price of the training and validation dataset between different models:

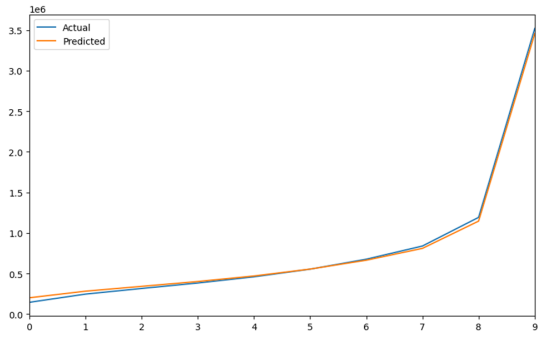
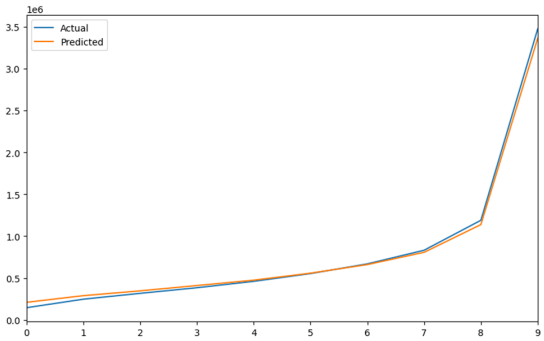
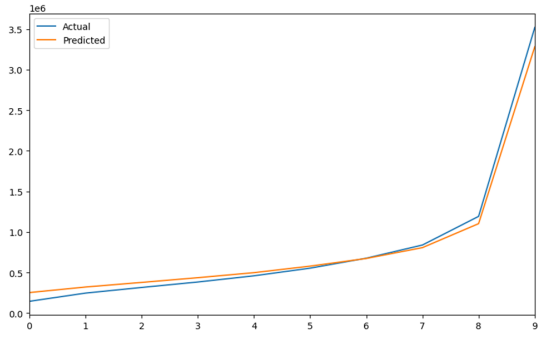
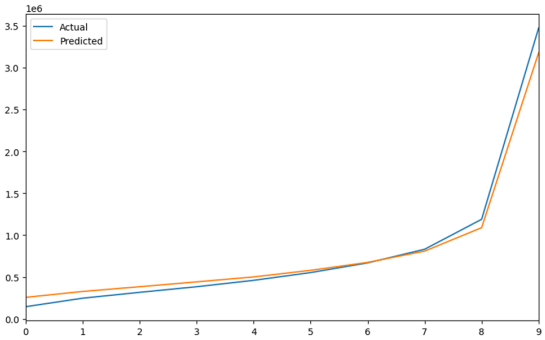
<table>
<tr><th></th><th>Training</th><th>Validation</th></tr>
<tr><td>RF</td><td>


<table>
<tr><th></th><th>bin</th><th>price</th><th>predicted_price</th><th>% diff</th></tr>
<tr><td>0</td><td>(20999.999, 209000.0]</td><td>1.460939e+05</td><td>1.575828e+05</td><td>7.864033</td></tr>
<tr><td>1</td><td>(209000.0, 280000.0]</td><td>2.473352e+05</td><td>2.571542e+05</td><td>3.969922</td></tr>
<tr><td>2</td><td>(280000.0, 349900.0]</td><td>3.168517e+05</td><td>3.275277e+05</td><td>3.369406</td></tr>
<tr><td>3</td><td>(349900.0, 419900.0]</td><td>3.840645e+05</td><td>3.938053e+05</td><td>2.536231</td></tr>
<tr><td>4</td><td>(419900.0, 499000.0]</td><td>4.603653e+05</td><td>4.694856e+05</td><td>1.981097</td></tr>
<tr><td>5</td><td>(499000.0, 600000.0]</td><td>5.546495e+05</td><td>5.625361e+05</td><td>1.421896</td></tr>
<tr><td>6</td><td>(600000.0, 749000.0]</td><td>6.773620e+05</td><td>6.789212e+05</td><td>0.230184</td></tr>
<tr><td>7</td><td>(749000.0, 949900.0]</td><td>8.396428e+05</td><td>8.366663e+05</td><td>-0.354492</td></tr>
<tr><td>8</td><td>(949900.0, 1499900.0]</td><td>1.192671e+06</td><td>1.182423e+06</td><td>-0.859293</td></tr>
<tr><td>9</td><td>(1499900.0, 875000000.0]</td><td>3.519178e+06</td><td>3.456506e+06</td><td>-1.780874</td></tr>
</table>
</td><td>


<table>
<tr><th></th><th>bin</th><th>price</th><th>predicted_price</th><th>% diff</th></tr>
<tr><td>0</td><td>(21899.999, 209000.0]</td><td>1.459123e+05</td><td>1.675648e+05</td><td>14.839408</td></tr>
<tr><td>1</td><td>(209000.0, 280000.0]</td><td>2.473823e+05</td><td>2.633098e+05</td><td>6.438426</td></tr>
<tr><td>2</td><td>(280000.0, 349900.0]</td><td>3.172891e+05</td><td>3.317388e+05</td><td>4.554100</td></tr>
<tr><td>3</td><td>(349900.0, 419900.0]</td><td>3.841754e+05</td><td>4.008875e+05</td><td>4.350137</td></tr>
<tr><td>4</td><td>(419900.0, 499000.0]</td><td>4.604341e+05</td><td>4.714851e+05</td><td>2.400128</td></tr>
<tr><td>5</td><td>(499000.0, 599900.0]</td><td>5.532866e+05</td><td>5.645708e+05</td><td>2.039484</td></tr>
<tr><td>6</td><td>(599900.0, 739000.0]</td><td>6.687597e+05</td><td>6.732590e+05</td><td>0.672783</td></tr>
<tr><td>7</td><td>(739000.0, 949000.0]</td><td>8.315160e+05</td><td>8.298930e+05</td><td>-0.195196</td></tr>
<tr><td>8</td><td>(949000.0, 1499000.0]</td><td>1.189142e+06</td><td>1.171227e+06</td><td>-1.506520</td></tr>
<tr><td>9</td><td>(1499000.0, 80000000.0]</td><td>3.471964e+06</td><td>3.361364e+06</td><td>-3.185527</td></tr>
</table>
</td></tr>
<tr><td>SVR</td><td>


</td><td>


</td></tr>
</table>

Light GBM _v1		
Light GBM _v2		

	bin	price	predicted_price	% diff
0	(20999.999, 209000.0]	1.460939e+05	2.434505e+06	1566.397176
1	(209000.0, 280000.0]	2.473352e+05	2.443545e+06	887.948948
2	(280000.0, 349900.0]	3.168517e+05	2.456867e+06	675.399743
3	(349900.0, 419900.0]	3.840645e+05	2.466726e+06	542.268691
4	(419900.0, 499000.0]	4.603653e+05	2.477145e+06	438.082372
5	(499000.0, 600000.0]	5.546495e+05	2.493036e+06	349.479564
6	(600000.0, 749000.0]	6.773620e+05	2.511117e+06	270.720023
7	(749000.0, 949900.0]	8.396428e+05	2.525771e+06	200.814999
8	(949900.0, 1499900.0]	1.192671e+06	2.551780e+06	113.955050
9	(1499900.0, 875000000.0]	3.519178e+06	2.597007e+06	-26.204166

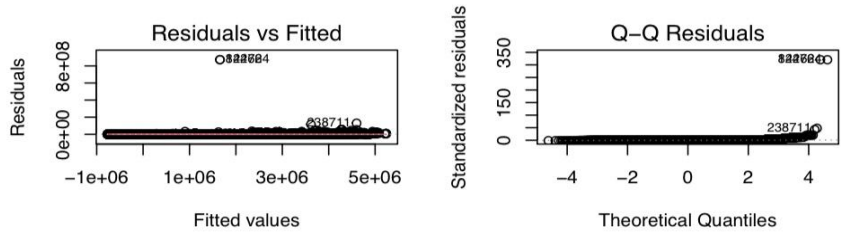
	bin	price	predicted_price	% diff
0	(21899.999, 209000.0]	1.459123e+05	400635.282593	174.572671
1	(209000.0, 280000.0]	2.473823e+05	420241.453090	69.875308
2	(280000.0, 349900.0]	3.172891e+05	464344.471567	46.347436
3	(349900.0, 419900.0]	3.841754e+05	506862.096714	31.935086
4	(419900.0, 499000.0]	4.604341e+05	534146.172706	16.009254
5	(499000.0, 599900.0]	5.532866e+05	558923.324264	1.018776
6	(599900.0, 739000.0]	6.687597e+05	585578.261992	-12.438168
7	(739000.0, 949000.0]	8.315160e+05	630584.674844	-24.164461
8	(949000.0, 1499000.0]	1.189142e+06	689984.785829	-41.976228
9	(1499000.0, 80000000.0]	3.471964e+06	876136.831705	-74.765386

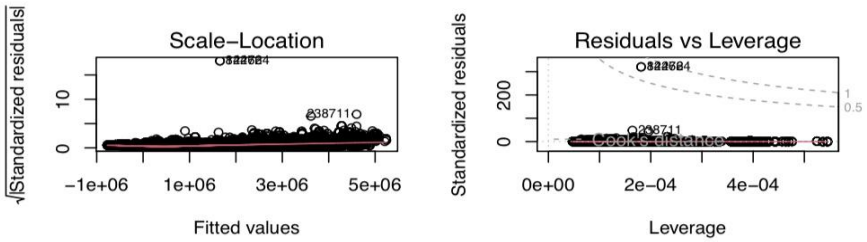
	bin	price	predicted_price	% diff
0	(20999.999, 209000.0]	1.460939e+05	2.019853e+05	38.257118
1	(209000.0, 280000.0]	2.473352e+05	2.833776e+05	14.572296
2	(280000.0, 349900.0]	3.168517e+05	3.428802e+05	8.214704
3	(349900.0, 419900.0]	3.840645e+05	4.025275e+05	4.807251
4	(419900.0, 499000.0]	4.603653e+05	4.708039e+05	2.267461
5	(499000.0, 600000.0]	5.546495e+05	5.555340e+05	0.159466
6	(600000.0, 749000.0]	6.773620e+05	6.647945e+05	-1.855363
7	(749000.0, 949900.0]	8.396428e+05	8.107328e+05	-3.443131
8	(949900.0, 1499900.0]	1.192671e+06	1.146647e+06	-3.858954
9	(1499900.0, 875000000.0]	3.519178e+06	3.457068e+06	-1.764898

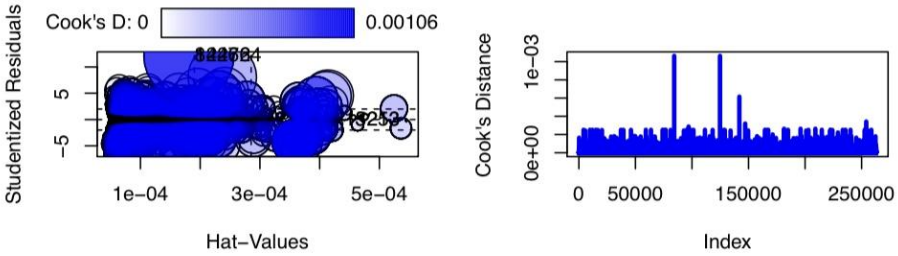
	bin	price	predicted_price	% diff
0	(21899.999, 209000.0]	1.459123e+05	2.101779e+05	44.043982
1	(209000.0, 280000.0]	2.473823e+05	2.894826e+05	17.018317
2	(280000.0, 349900.0]	3.172891e+05	3.468370e+05	9.312618
3	(349900.0, 419900.0]	3.841754e+05	4.104689e+05	6.844137
4	(419900.0, 499000.0]	4.604341e+05	4.744661e+05	3.047552
5	(499000.0, 599900.0]	5.532866e+05	5.580507e+05	0.861051
6	(599900.0, 739000.0]	6.687597e+05	6.603861e+05	-1.252109
7	(739000.0, 949000.0]	8.315160e+05	8.070221e+05	-2.945697
8	(949000.0, 1499000.0]	1.189142e+06	1.137859e+06	-4.312585
9	(1499000.0, 80000000.0]	3.471964e+06	3.359834e+06	-3.229606

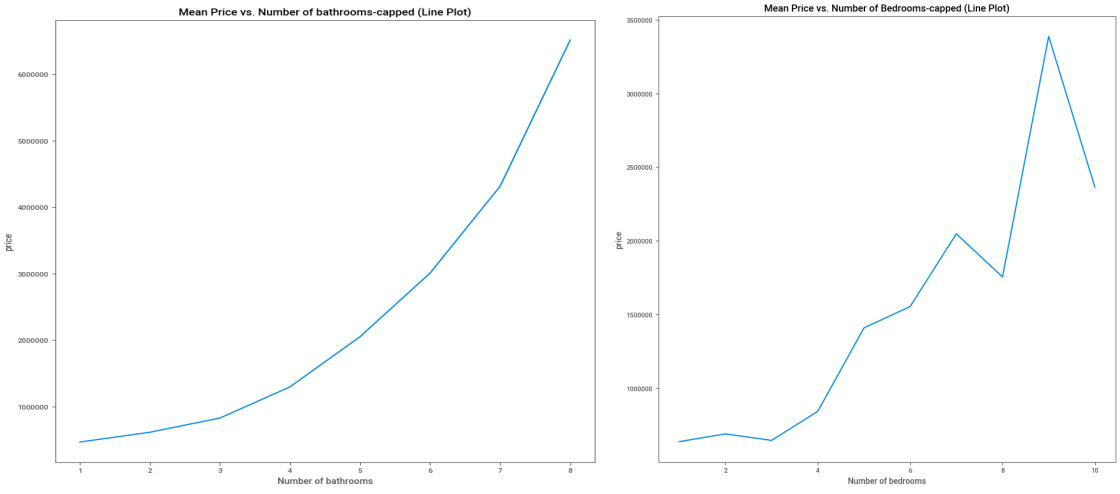
	bin	price	predicted_price	% diff
0	(20999.999, 209000.0]	1.460939e+05	2.538434e+05	73.753522
1	(209000.0, 280000.0]	2.473352e+05	3.221736e+05	30.257923
2	(280000.0, 349900.0]	3.168517e+05	3.790617e+05	19.633767
3	(349900.0, 419900.0]	3.840645e+05	4.371439e+05	13.820441
4	(419900.0, 499000.0]	4.603653e+05	4.991920e+05	8.433888
5	(499000.0, 600000.0]	5.546495e+05	5.796385e+05	4.505362
6	(600000.0, 749000.0]	6.773620e+05	6.739745e+05	-0.500112
7	(749000.0, 949900.0]	8.396428e+05	8.075618e+05	-3.820788
8	(949900.0, 1499900.0]	1.192671e+06	1.102976e+06	-7.520549
9	(1499900.0, 875000000.0]	3.519178e+06	3.281978e+06	-6.740206

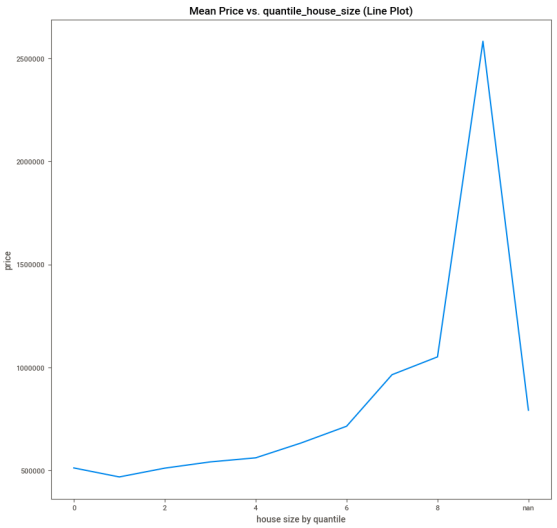
	bin	price	predicted_price	% diff
0	(21899.999, 209000.0]	1.459123e+05	2.572669e+05	76.316124
1	(209000.0, 280000.0]	2.473823e+05	3.278555e+05	32.529904
2	(280000.0, 349900.0]	3.172891e+05	3.842100e+05	21.091472
3	(349900.0, 419900.0]	3.841754e+05	4.423167e+05	15.134049
4	(419900.0, 499000.0]	4.604341e+05	5.011063e+05	8.833448
5	(499000.0, 599900.0]	5.532866e+05	5.800385e+05	4.835102
6	(599900.0, 739000.0]	6.687597e+05	6.747459e+05	0.895112
7	(739000.0, 949000.0]	8.315160e+05	8.091364e+05	-2.691427
8	(949000.0, 1499000.0]	1.189142e+06	1.089163e+06	-8.407584
9	(1499000.0, 80000000.0]	3.471964e+06	3.176580e+06	-8.507718

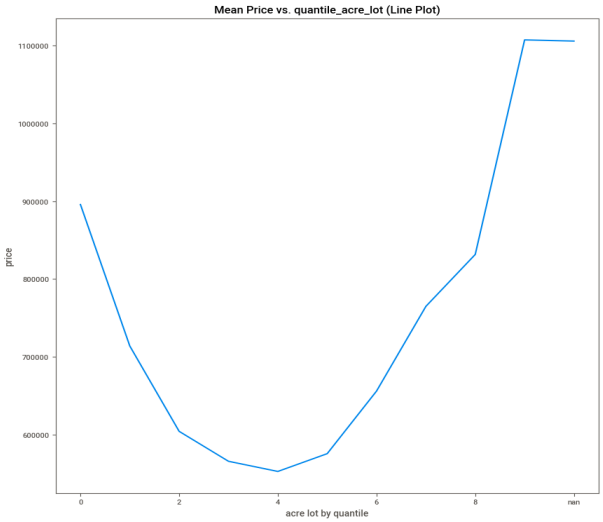

Residuals vs Fitted
Residuals
Fitted values
Q–Q Residuals
Standardized residuals
Theoretical Quantiles

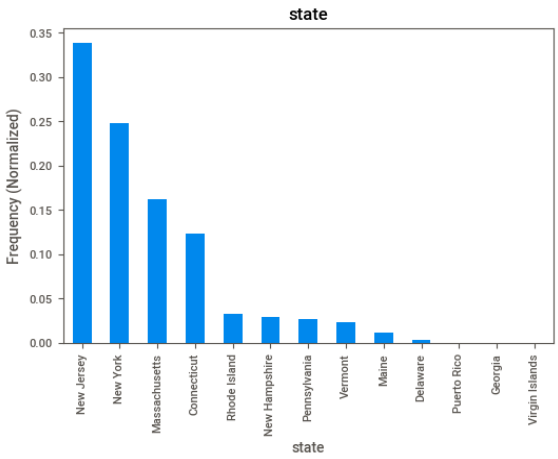

Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage


Cook's D: 0
0.00106
Studentized Residuals
Hat–Values
Cook's Distance
Index


Mean Price vs. Number of bathrooms-capped (Line Plot)
price
Number of bathrooms
Mean Price vs. Number of Bedrooms-capped (Line Plot)
price
Number of bedrooms


Mean Price vs. quantile_house_size (Line Plot)
price
house size by quantile


Mean Price vs. quantile_acre_lot (Line Plot)
price
acre lot by quantile


state
Frequency (Normalized)
New Jersey
New York
Massachusetts
Connecticut
Rhode Island
New Hampshire
Pennsylvania
Vermont
Maine
Delaware
Puerto Rico
Georgia
Virgin Islands
state