Analysis Plan - Housing Prices in the USA

Michael Butros

Due: April 27, 2025

1 Background

Housing prices in the United States are influenced by multiple factors, including economic conditions, demographic trends, interest rates, and government policies. Understanding these factors can help policymakers, investors, and homebuyers make informed decisions. Over the years, housing markets have exhibited cyclical trends, experiencing booms and busts due to economic fluctuations, lending policies, and housing supply-demand dynamics. Analyzing these factors can provide insights into market behavior and potential future trends.

The House Price Index (HPI) is frequently utilized to gauge fluctuations in housing prices. However, because housing prices are heavily influenced by factors such as location, area, and population, additional information beyond the HPI is necessary to accurately predict individual housing prices.

Numerous studies have employed traditional machine learning techniques to forecast housing prices with precision. However, these studies often overlook the performance of individual models and tend to ignore less popular but more complex models. Consequently, to examine the various impacts of different features on prediction methods. This section of the project will employ a variety of machine learning techniques to explore the differences among several models in predicting the HPI.

Using a Kaggle dataset compiled from publicly available sources, factors that might affect the price of homes in the US were analyzed using machine learning techniques to ascertain the effect of these factors on the house price index.

Techniques to be used include: linear and generalized regression, lasso regression, elastic nets, k-nearest neighbors, classification and regression trees, and support vector machine.

2 Literature Review

Recent advances in housing price prediction have been heavily influenced by machine learning (ML) methodologies, particularly ensemble models. Random Forest and XGBoost have consistently demonstrated high accuracy in modeling nonlinear patterns in real estate data. Sharma et al. (2024) and Tanamal et al. (2023) both found that Random Forest provided robust predictive results when applied to housing data in different contexts, with accuracies exceeding 85%. Similarly, the XGBoost algorithm has emerged as a leading choice for regression-based tasks. Wu (2024) proposed a hybrid model combining XGBoost with Bagging to reduce variance while maintaining bias reduction benefits, outperforming standalone models. These results are consistent with other ML-based studies, such as Kuvalekar et al. (2020), who achieved 89% accuracy using a Decision Tree Regressor model incorporating both structural and environmental features. The adoption of advanced models aligns with Zhang (2021) and Liu (2022), who demonstrated that although multiple linear regression (MLR) is interpretable and effective in some settings, more complex models often yield better accuracy in volatile or high-dimensional data environments.

Behavioral factors also play a crucial role in housing price dynamics, particularly regarding how list prices are set and adjusted. Hayunga and Pace (2016) demonstrated that seller characteristics—including income, expectations of loss, and urgency—influence initial list prices and subsequent adjustments. Their findings support prospect theory, particularly the concept of loss aversion, where anticipated losses prompt sellers to list properties above expected market value. Similarly, Haurin et al. (2010) found that atypical properties are often overpriced initially and remain on the market longer, underscoring how strategic pricing behavior affects transaction timelines and outcomes. These insights reinforce the need for predictive models to account not just for property features but also seller-side dynamics.

Housing market synchronization and regional spillover effects have also been studied extensively. Gupta et al. (2021) used Bayesian dynamic factor models and random forests to show that macroeconomic uncertainty significantly predicts synchronization across U.S. state housing markets. Their findings indicate that national-level shocks can explain substantial portions of regional housing price variance. This complements earlier work by Moench and Ng (2011), who found that national housing shocks have persistent effects on retail consumption across U.S. regions. Yunus and Swanson (2013) extended this understanding by showing increased cointegration among regional housing markets post-2006, suggesting reduced diversification benefits and increased systemic risk. Antonakakis et al.

(2021) similarly confirmed dynamic spillovers across regions, particularly during economic downturns, with sales volume often acting as a leading indicator.

Finally, the policy and theoretical foundations of housing market analysis have evolved alongside empirical methods. The Congressional Research Service (2023) highlighted the persistent mismatch between housing supply and demand in the U.S., pointing to zoning regulations, labor shortages, and pandemic-related disruptions as key constraints. These reports emphasize the importance of integrating structural and policy variables into predictive models, especially as housing affordability becomes an increasingly salient public policy concern.

In summary, the literature supports the growing dominance of ensemble and hybrid machine learning models in housing price prediction. While traditional regression remains relevant for its transparency, models like Random Forest, XGBoost, and stacked ensembles provide superior performance, especially when coupled with robust preprocessing and hyperparameter tuning. The field is trending toward more dynamic, hybridized systems that incorporate spatial, temporal, and economic complexity in real estate valuation.

The most common evaluation metrics across the literature include RMSE, MAE, MAPE, R^2, and RMSLE. Cross-validation techniques, especially k-fold, are standard practices to mitigate overfitting. Ritu (2023) highlighted that selecting the appropriate evaluation metric is critical and often dataset-specific, depending on skewness, scale, and domain objectives.

Despite significant progress, challenges remain. Many studies neglect macroeconomic and temporal factors (e.g., inflation, interest rates), limiting model generalizability. The scalability of stacked models is also limited by their computational cost. Moreover, spatial heterogeneity and real-time adaptability are often underaddressed.

3 Exploratory Data Analysis (EDA)

The dataset for the HPI factors for this analysis section was obtained from Kaggle. The dataset is found in the file *US House Price Factors* and was downloaded as a .csv file. It was compiled from publicly available sources, with at least three sources of data: GDP, interest rates, and mortgage rates, among others.. The dataset contained factors that are not commonly used in the literature review for the project. The dataset variables consist of only numeric data except for a monthly date tracker.

The numeric variables are described below:

- PREDICTORS:
  - Building Permit: Examine new constructions
  - Construction Price Index: Look at trends in construction costs
  - Delinquency Rate: Consider consumer payment behaviors
  - GDP: How the national economy impacts housing (Gross Domestic Product)
  - House Sold or For Sale: Track market dynamics
  - Housing Subsidies: Understand government interventions
  - Income: Consider the financial backbone of homeownership
  - Interest Rate: Realize the cost of borrowing
  - Mortgage Rate: Look at the shape of financing landscape
  - New Construction Units: Include new house developments
  - Total Houses: Look at the housing availability
  - Total Construction Spending: Consider financial trends of the construction sector spending
  - Unemployment Rate: Examine the effect of unemployment's percentage
  - Urban Population: Spotlight on city living
- TARGET VARIABLE:
  - Home Price Index (HPI)

Note that the dataset contains a wide range of factors covering: **economic factors**: GDP., income, unemployment rate, $\cdots$, **housing factors** : building permits, house prices, interest/mortgage rates, $\cdots$; **construction factors**: new construction units, total construction spending, $\cdots$. Also note that the dataset does not include the common factors we think of when purchasing a home. The common factor usually used in this type of analysis include, but are not limited to: number of bedrooms, number of bathrooms, square footage of the house, attached garage, parking spots, $\cdots$.

The dataset spans over 20 years, providing enough time to detect trends across:

- Pre-2008 housing crisis

- Post-2008 housing recovery
- COVID-19 pandemic effects (2020-2021)
- Recent inflation/interest rate adjustments (2022-2023)

A print out of the sum of missing values for each column in the dataset revealed that the dataset did not contain any missing values, which indicates a clean dataset, in regards to missing values. Next, we look at the data and obtain a summary of each of the columns.

Statistical Summary Key Findings:

- The home price index shows a substantial spread, indicating high variability in housing prices over time.
- The delinquency rate also shows wide variation, peaking above 11%, suggesting periods of financial stress.
- Construction spending includes negative values, implying possible contractions or revisions.
- Urban population is highly stable, clustered around 81%..

We draw the histogram for each numerical variable in the dataset to obtain an idea about data distribution and to initially determine the normality or skewness of the variables.

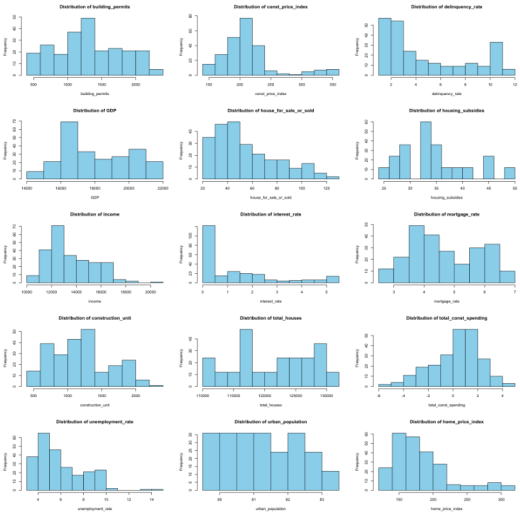

Figure 1: Variable Distribution

Here's a summary of the visual findings (see Figure 1):

1. Skewed Distributions: Several variables are right-skewed (positively skewed):
    - Delinquency Rate: Most values are low; a few high outliers.
    - Interest Rate: Majority below 2%; a few much higher.
    - Unemployment Rate: Mostly clustered at lower rates.
    - Total Construction Spending: Mostly below 2; with a noticeable tail.
    - Home Price Index: Majority values between 130–200, but extends to above 300.
2. Approximately Normal or Symmetric Distributions: These variables exhibit a more balanced bell-shaped or uniform-like distribution:
    - Building Permits: Fairly symmetric but with some dispersion.
    - GDP: Roughly bell-shaped around the median ($\sim$ 18,000).
    - Income: Skewed slightly right but closer to normal than others.
    - Mortgage Rate: Nearly symmetric, slight right skew.
3. Bimodal or Multimodal Distributions
    - Housing Subsidies: Two peaks suggest bimodality, possibly due to policy shifts or changes in administration.
    - Construction Units: May indicate multiple activity levels or cycles.
4. Uniform or Narrow Range
    - Urban Population: Very tightly clustered (between 79.5 and 83); suggests minimal variability.
    - Total Houses: Distribution is compact and symmetric.
5. Other Notables
    - House for Sale or Sold: Slight right skew; most values under 80.
    - Construction Price Index: Mostly between 180–240, with a long tail on the higher end.

Next, we examined the correlation matrix between the variables. The correlation matrix can be used for identifying relationships between variables, in feature selection for modeling, data reduction and simplification, and data cleaning and

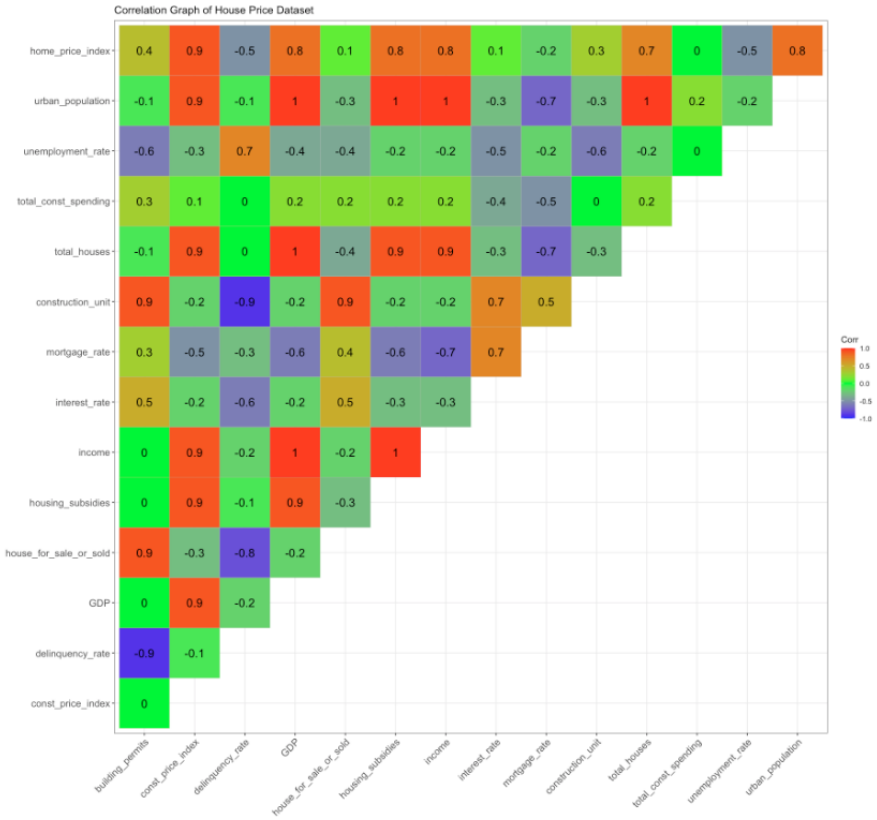

Figure 2: Correlation Matrix

validation.

Based on the correlation matrix, we find the following (see Figure 2) (not an exhaustive list):

- A positive correlation exists between home price index and construction price index, GDP, housing subsidies, income, total houses, and urban population. This means that home price index is strongly associated with higher values of these variables.

- A negative correlation exists between home price index and delinquency rate. This means as home prices rise, delinquency rates fall - likely due to better economic conditions. Also, between construction unit and delinquency rate. Finally, between unemployment rate and each of income and GDP. This means that higher unemployment is associated with lower income and GDP, as expected.

- A positive correlation exits between mortgage rates and interest rates. This is expected since interest rates drive mortgage rates.

Finally, we examine the relation between some of the variables in the dataset and their changes over time . Plots for all the variables were computed, however, we only mention a few findings here:

- Income: As we would expect (hope) income over the past 20 years of the dataset shows a generally increasing trend until the year 2021-2022.

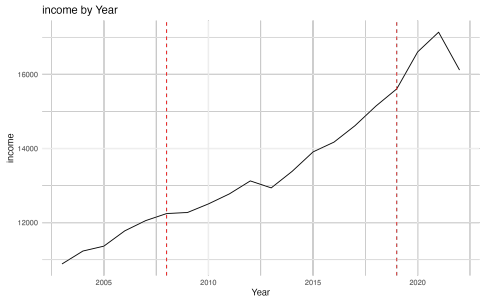

Figure 3: Income vs Time

- Interest Rate: The trends show a decrease in interest rates during the housing market class between 2008-2010 with minimal changes in rates until 2015 when rates began to increase steadily until the COVID-19 pandemic. Rates decreased after the pandemic until 2021-2022 when they began to increase again.

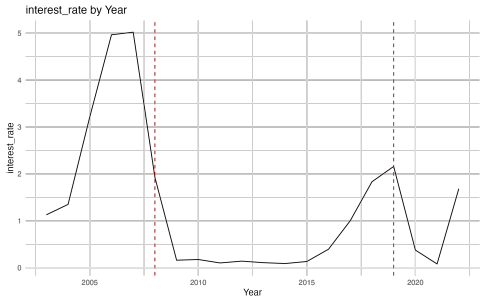

Figure 4: Interest Rate vs Time

- Gross Domestic Product (GDP): The GDP shows a growth shortly after the housing market clash of 2008 until the beginning of the COVID-19 pandemic. when it slipped for a year and then began rise again.

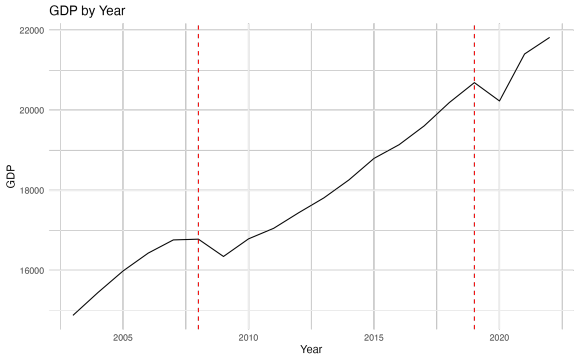

Figure 5: GDP vs Time

- Delinquency Rate: We can easily see the high delinquency rate during the housing market crash that began in 2008 reaching the highest point in 2010 before the recovery of the market began and the delinquency rate beginning to decrease.

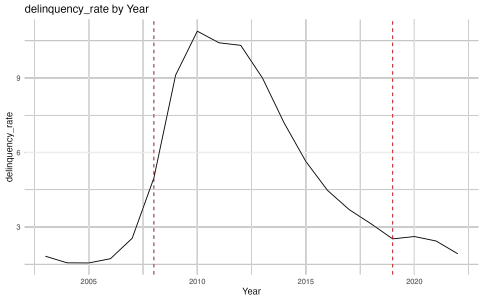

Figure 6: Delinquency Rate vs Time

- Home Price Index: We notice a decrease in the home price index between 2007 and 2012. Home price index seem to be on a steady increase since then with a steep slope after the COVD-19 pandemic.

Key EDA Findings:

- Income, Gross Domestic Product (GDP) and employment are strong drives of home prices.
- Years 2008 - Housing Crisis and 2019-2020 of the COVID-19 Pandemic stand out as economic shock periods, disturbing trends.
- Even though we are not doing this for this project, the dataset is suitable for time series modeling and forecasting home prices.

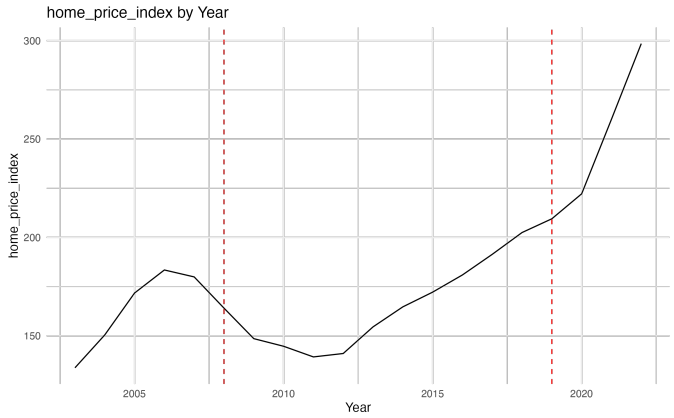

Figure 7: Home Price Index vs Time

4 Regression Analysis and Variable Selection

In this section we perform some regression techniques on the housing dataset and evaluate the performance of the different models.

4.1 Linear Regression: Full and Reduced Models

We begin our regression analysis with a full model for home price index. The full model contains the HPI as the response variable and all other variables as potential significant coefficients. The summary of the full model shows that some statistically significant factors do exist. These factors are: construction price index, delinquency rate, housing subsidies, income, interest rate, construction units, and total houses. We then ran a model with the statistically significant coefficients as a reduced model.

Next we checked the assumptions for regression by getting residual plots and The reduced model (see Figures 8 and 9). In Figure 8, the residuals seem randomly scattered around zero but with some curved patterns (non-linearity), and variance seems to increase with fitted values (possible heteroscedasticity). This is a concern since the curvature and increasing spread may suggest the model does not fully capture the underlying structure. While Figure 9 suggests some deviation from normality but not severe.

We can therefore conclude the following about the assumptions:

- Residuals versus fitted values show some curvature suggesting a non-linear relationship.

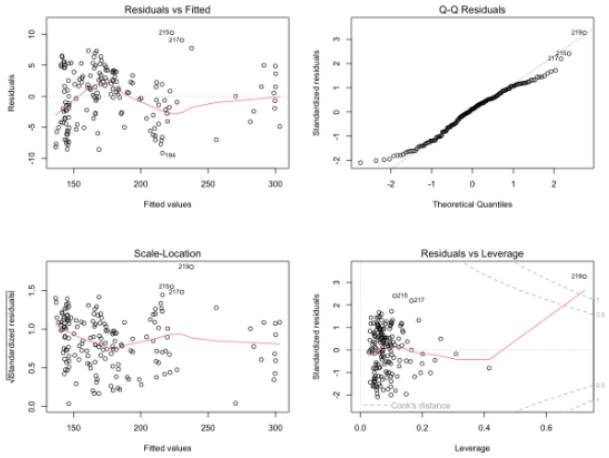

Figure 8: Assumptions for Regression Modeling: Residual Plots

- Scale-Location and Residuals versus Fitted values show increasing spread suggesting possible heteroscedasticity.
- The Residuals versus Leverage plot suggests the existence of influential points (outliers).
- There are small deviations in the tail of the QQ-Plot but points mostly follow the line suggesting a normality.

The above conclusions suggested we perform a variance inflation factor (VIF) analysis on the variables to measure how strongly predictable variables are linearly related to others. The analysis suggested the existence of highly related predictor variables (high VIF value) and some with moderate to low VIF values (Values between 10 and VIF threshold).

Based on the VIF results we ran a reduced model using the predictor variables below. The selected variables were all with moderate to low VIF value and/or that were statistically significant from the full model.

Reduced Model Predictors:

- Construction Price Index
- Delinquency Rate
- Housing Subsidies
- Construction Unit

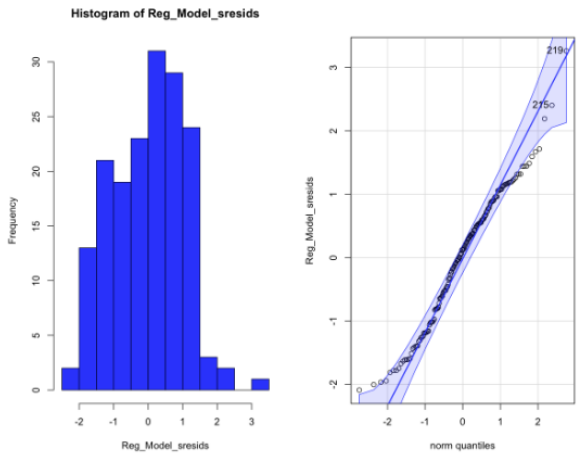

Figure 9: Assumptions for Regression Modeling: Histogram and QQ-Plot

- Income
- Unemployment Rate
- Mortgage Rate
- Interest Rate

A comparison of model summaries between the full and reduced models yielded the following table:

Table 1: Full versus Reduced Regression Model

Model	MSE	df	R^2	Adjusted R^2	F-Statistic	df	p-value
Full	4.447	224	0.9891	0.9884	1356	15 and 224	2.2×10^{-16}
Reduced	4.8	231	0.9869	0.9865	2178	8 and 231	2.2×10^{-16}

Next we conducted an ANOVA test between the full and reduced model. The ANOVA test yielded an F value of 6.43652 and a p-value of 6.4847×10^{-7}. Both of the F value and p-value indicate that the subset of variables (those in the reduced model) contain valuable information about the variability of the HPI.

4.2 Stepwise regression: Forward, Backward, and Both Directions

To identify the most significant predictors for the response variable, a stepwise regression approach was employed. This iterative model selection method adds

or removes predictors based on specific criteria (typically AIC, BIC, or p-values) to achieve a balance between model complexity and explanatory power. The goal is to construct a model that retains only variables contributing meaningfully to the prediction, while minimizing overfitting. Both forward selection and backward elimination and both directions were considered in the analysis.

We first applied a forward stepwise regression starting with a model that only contained the intercept. The starting model had an AIC of 1786.5. The final model, for this approach, had an AIC of 755.75 and the following coefficients:

Table 2: Stepwise Forward Final Model Coefficients

Coefficient	Value
Intercept	-38.9407
construction price index	0.632080
delinquency rate	-2.68029
interest rate	4.05394
unemployment rate	1.26357
housing subsidies	1.84256
construction unit	0.02161
house for sale or sold	-0.08295

Next, we applied a backward stepwise regression starting with a full model. The starting model had an AIC of 731.76. The final model, for this approach, had an AIC of 727.8 and the following coefficients:

Table 3: Stepwise Backward Final Model Coefficients

Coefficient	Value
Intercept	-5.266×10^{2}
Date	1.089×10^{-2}
construction price index	5.998×10^{-1}
delinquency rate	-2.889×10^{0}
housing subsidies	1.411×10^{0}
income	-2.599×10^{-3}
interest rate	4.401×10^{0}
mortgage rate	1.008×10^{0}
construction unit	1.429×10^{-2}
total houses	-4.183×10^{-3}
unemployment rate	1.054×10^{0}
urban population	1.086×10^{1}

NOTE: The model above contain two coefficients which raise a flag. Those coefficients are Date and Urban population because they had the highest two VIF values in the dataset. We will deal with multicollinearity in the next subsection.

We next performed a stepwise regression in both directions. We did this approach once beginning with a full model and once beginning with a minimum model (containing only the intercept). The model starting with the minimum model began with an AIC of 1786.5 and the final model had an AIC of 755.75 and the following coefficients:

Table 4: Stepwise Forward and Backward - Minimum Start Final Model Coefficients

Coefficient	Value
Intercept	-38.94070
construction price index	0.632080
delinquency rate	-2.680290
interest rate	4.053940
unemployment rate	1.263570
housing subsidies	1.842560
construction unit	0.021610
house for sale or sold	-0.082950

NOTE: this is the same model we obtained when the stepwise forward only direction was used.

The model starting with the full model began with an AIC of 731.76 and the final model had an AIC of 727.8 and the following coefficients:

Table 5: Stepwise Backward Final Model Coefficients

Coefficient	Value
Intercept	-5.266×10^{2}
Date	1.089×10^{-2}
construction price index	5.998×10^{-1}
delinquency rate	-2.889×10^{0}
housing subsidies	1.411×10^{0}
income	-2.599×10^{-3}
interest rate	4.401×10^{0}
mortgage rate	1.008×10^{0}
construction unit	1.429×10^{-2}
total houses	-4.183×10^{-3}
unemployment rate	1.054×10^{0}
urban population	1.086×10^{1}

NOTE: this is the same model we obtained when the stepwise backward only direction was used. Also note that the model contains the Date and Urban population coefficients which had the highest two VIF values in the dataset.

4.3 Ridge, Lasso, Elastic Net, Group Lasso, and Random Forest

To address issues of multicollinearity, overfitting, and variable selection, a suite of regularized regression techniques was applied, including Ridge, Lasso, Elastic Net, and Group Lasso. These methods impose penalties on the magnitude of regression coefficients to enhance model generalizability and interpretability. Ridge regression applies an L2 penalty to shrink coefficients, Lasso employs an L1 penalty to encourage sparsity, and Elastic Net combines both penalties to balance their effects. Group Lasso extends this concept by selecting or discarding entire groups of related variables simultaneously.

Additionally, Random Forest, an ensemble learning method based on decision trees, was used to capture non-linear relationships and complex interactions among predictors. It serves as a powerful benchmark due to its robustness to overfitting and ability to rank variable importance.

We began by running a ridge regression model using *lm.ridge()* on the predictors in the dataset, without the date variable. We also applied a range of penalty constants between 0 and 10 by a step of 0.2 that resulted in the following model coefficients:

Table 6: Ridge Regression Model Coefficients

Coefficient	Value
building permits	0.0681
construction price index	0.5880
delinquency rate	-0.1835
GDP	0.2114
house for sale or sold	-0.0479
housing subsidies	0.2538
income	-0.0316
interest rate	0.1359
mortgage rate	0.0584
construction unit	0.1715
total houses	-0.1354
total construction spending	-0.0068
unemployment rate	0.0894
urban population	0.0910

We then ran a lasso regression model using *glmnet()* on the same predictors we used for the ridge regression model. We use an α value of 1 and 100 λ's. We found the optimal λ using a 10-fold cross validation. For this model, we obtained the following coefficients for the optimal λ:

Table 7: Lasso Regression Model Coefficients

Coefficient	Value
Intercept	-1.975018×10^{1}
building permits	2.34977×10^{-3}
construction price index	5.8654585×10^{-1}
delinquency rate	-2.578636×10^{0}
GDP	4.685331×10^{-3}
house for sale or sold	-5.676172×10^{-2}
housing subsidies	1.715577×10^{0}
income	-6.615087×10^{-4}
interest rate	3.633862×10^{0}
mortgage rate	1.399869×10^{0}
construction unit	1.73761×10^{-2}
total houses	-7.45731×10^{-4}
total construction spending	0.0
unemployment rate	1.86088×10^{0}
urban population	0.0

For the elastic net regression model, we ran an α value of 0.5. We again fit a

model with 100 λ's and performed a 10-fold cross validation. We then extracted the following optimal λ coefficients:

Table 8: Elastic Net Regression Model Coefficients

Coefficient	Value
Intercept	-51.365884835
building permits	0.002169119
construction price index	0.596743947
delinquency rate	-2.393837580
GDP	0.002860913
house for sale or sold	-0.040777756
housing subsidies	1.630493372
income	0.0
interest rate	3.644705975
mortgage rate	1.598854294
construction unit	0.017968950
total houses	-0.000291137
total construction spending	0.0
unemployment rate	1.546447082
urban population	0.0

Next, we conducted a group lasso regression. We scaled the dataset variables, both predictors and response. Came up with a group assignment and performed a 10-fold cross validation for the group lasso using *cv.gglasso()* with 100 λ's to obtain the following model coefficients:

Table 9: Group Lasso Regression Model Coefficients

Coefficient	Value
Intercept	2.577375×10^{-15}
building permits	5.059250×10^{-2}
construction price index	6.232516×10^{-1}
delinquency rate	-2.068580×10^{-1}
GDP	2.307581×10^{-1}
house for sale or sold	-4.523166×10^{-2}
housing subsidies	2.405549×10^{-1}
income	-4.510100×10^{-2}
interest rate	1.409854×10^{-1}
mortgage rate	4.336154×10^{-2}
construction unit	1.682834×10^{-1}
total houses	-1.356172×10^{-1}
total construction spending	-3.023749×10^{-3}
unemployment rate	9.853533×10^{-2}
urban population	5.743575×10^{-2}

5 Model Performance

To evaluate the performance of various regression models, we analyze and compare key metrics such as Root Mean Squared Error (RMSE) and R^2 across multiple algorithms. This includes traditional Linear Regression, which assumes a linear relationship between predictors and the target; Lasso and Ridge Regression, which introduce regularization to handle multicollinearity and prevent overfitting; and ensemble methods like Random Forest and Gradient Boosting Machines (GBM), which are capable of capturing complex, nonlinear patterns in the data. By examining their performance on the same dataset, we can assess each model's ability to generalize and identify the most suitable approach for the regression task at hand.

Applying the evaluation metrics above we obtain the following table for comparison between different techniques:

Table 10: Performance Metrics of Different Regression Models

Model	RMSE	R^2
Linear	4.957248	0.9874148
Lasso	4.957800	0.9874113
Ridge	6.225747	0.9801488
Random Forest	2.47230	0.9970075
Gradient Boosting	3.646352	0.9931904

Key findings from the performance table

- Random Forest has the best performance, showing the lowest RMSE (2.47230) and highest R^2 (0.9970075), indicating highly accurate predictions and excellent fit to the data.
- Gradient Boosting also performs well, with relatively low RMSE (3.646352) and high R^2 (0.9931904), slightly behind Random Forest.
- Linear and Lasso regression perform almost identically, suggesting that regularization in Lasso does not significantly improve performance for this dataset.
- Ridge regression performs worse than both Linear and Lasso, with the highest RMSE and lowest R^2 , indicating a weaker fit.
- Ensemble models (Random Forest and Gradient Boosting) outperform linear models in this case, capturing more complexity in the data, while regularization (Lasso and Ridge) doesn't notably improve over simple linear regression here.

To optimize the performance of the regression models parameter tuning is essential. For Lasso and Ridge regression, tuning involves selecting the optimal regularization parameter alpha using techniques like cross-validated grid search over a range of alpha values. For ensemble methods such as Random Forest and Gradient Boosting, tuning focuses on hyperparameters such as the number of estimators, maximum tree depth, learning rate (for boosting), and minimum samples per leaf. These parameters can be optimized using grid search or randomized search in combination with cross-validation to avoid overfitting and ensure model generalization. Implementing this systematic tuning process helps identify the most effective model configuration for achieving lower error rates and higher predictive accuracy.

We now present the best results obtained from training and tuning different regression models:

- Linear Cross Validation (method = "lm") with $k = 10$

Table 11: Linear Model Cross Validation

Intercept	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
True	5.725288	0.9700222	4.438698	2.401935	0.04161168	1.427841

- Lasso Cross Validation (method = "glmnet") with $k = 10$

Table 12: Lasso Model Cross Validation

α	λ	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
1	0.06981762	5.377564	0.9807248	4.290728	0.5958012	0.01126912	0.3770712

- KNN Cross Validation (method = "knn") Different K values

  We tried $k = 10$ but it did not results with the best values so we tested over the following k values: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21.

Table 13: KNN Model Cross Validation

k	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
1	2.426792	0.9965382	1.794926	0.6635952	0.001663801	0.3756727

- Elastic Net (method = "ranger") Minimum Node Size of 5, number of tree of 10

Table 14: Elastic Net Model Cross Validation

$mtry$	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
3	3.300766	0.9936751	2.336178	1.048781	1.048781	0.6146201

- Classification and Regression Tree (method = "rpart") with $k = 10$

Table 15: CART Model Cross Validation

cp	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
0.1118157	12.10021	0.9079163	9.969641	3.253075	0.08794765	2.797231

- Support Vector Machine (method = "svmRadial") with $k = 10$

Table 16: SVM Model Cross Validation

σ	C	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
0.3438073	1.00	6.281557	0.9738666	3.789179	3.898701	0.03556756	1.150067

- Random Forest (method = "rf") with 10 folds ($ntrees = 10$)

Table 17: Random Forest Model Cross Validation

$mtry$	$RMSE$	R^2	MAE	$RMSE_{SD}$	R^2_{SD}	MAE_{SD}
3	3.610411	0.9940965	2.302224	1.538499	0.004391177	0.6895934

- We also tried to train and tune a PCA model but it resulted in the highest value for RMSE (16.47035) and lowest R^2 value (0.8407625).

6 Conclusions

6.1 Challenges:

Using all numeric predictor variables in our regression analysis offers some advantages—simplicity, ease of preprocessing, and compatibility with a wide range of models—but it also presents several challenges that may affect model performance and interpretability, especially in the context of the Home Price Index.

Challenges of Using All Numeric Predictors in Regression Analysis:

- Loss of Informational Granularity: If categorical features (e.g., location, zoning type, housing style) are excluded or one-hot encoded into numeric form, important context or nonlinear relationships might be diluted or lost, potentially limiting the model's predictive power.
- Multicollinearity Among Predictors:Numeric variables often correlate with each other (e.g., square footage and number of rooms), leading to multicollinearity, which inflates variance in coefficient estimates for linear models (Linear, Lasso, Ridge), making interpretations unstable.
- Scaling and Distributional Issues: Models like Lasso, Ridge, and Gradient Boosting can be sensitive to differences in scale. Variables with larger magnitudes can dominate the learning process unless proper normalization or standardization is applied.
- Nonlinearity and Interaction Effects: Pure numeric input assumes linear or additive effects unless explicitly modeled. While ensemble methods like Random Forest and Gradient Boosting can capture nonlinearities and interactions, linear models cannot unless interaction terms are manually added.
- Interpretability Tradeoffs: As the model complexity increases (e.g., with ensembles), even though all variables are numeric, it becomes more difficult to interpret individual feature importance, particularly when numeric predictors interact in nonlinear ways.
- Overfitting Risks in Flexible Models: With many numeric variables, especially if they are noisy or redundant, models like Random Forest and Gradient Boosting may overfit without careful tuning and regularization, despite their superior performance.

6.2 Summary

In this regression analysis project on the Home Price Index, a variety of regression models—including Linear Regression, Lasso, Ridge, Random Forest, and Gradient Boosting—were employed to evaluate predictive performance. Through a rigorous training and testing framework, each model was assessed using key metrics such as RMSE and R^2. Furthermore, hyperparameter tuning was conducted to optimize model performance, particularly for the more flexible models like Random Forest and Gradient Boosting. Among the models tested, ensemble methods such as Random Forest and Gradient Boosting demonstrated superior predictive accuracy, outperforming traditional linear approaches in capturing the underlying patterns in the data. Overall, this analysis highlights the importance of model selection and tuning in regression tasks, and confirms that ensemble learning techniques are particularly effective for modeling complex, non-linear relationships in housing market data.