Housing Price Prediction

Jennifer Dong
jd1358
Section 5

Ryan Bae
jb1995
Section 7

Joshua Gohil
jg1785
Section 6

Abstract

This study analyzes the changes in housing prices across the United States over a five-year span, with particular attention to increases following the COVID-19 pandemic. Using housing market data from Redfin, our goal was to develop predictive models to forecast housing prices based on region-specific features. By applying supervised machine learning techniques, we evaluated their effectiveness in predicting future housing market trends. Our findings highlight the differences in model performance based on geographic factors and explores the complexities involved in accurately forecasting real estate prices.

1 Introduction

Problem Statement

The problem we set out to solve was the prediction of house prices. Since the market for housing has been rising ever since the COVID-19 Pandemic, we aimed to not only create a project that can determine the price of a house given the region, previous pricing (monthly range), inventory, and other features, but also a good predictor of if trends ever showed a decrease in prices. While price prediction is the main goal our group is trying to solve, predicting future trends is also another crucial component of our project.

Plan

Using a supervised learning model, our group aimed to use a dataset collected from a well-known real estate company, Redfin, and create a predictive model. After preprocessing the data, we will then feed it into two machine learning models: linear regression and random forest, which will be able to predict the median of a region's house prices given the year, month, type of residence, inventory, etc. After creating efficient models, we shall test and improve them depending on how accurate they are. Afterwards, our group plans to visualize the data so it can be interpreted and analyzed.

2 Motivation

Importance

The idea of predicting house prices is very important to statisticians and economists, who can use the data to predict pricing trends and, in general, how the economy will perform. Housing is a staple of the American dream, so its economic impact is strong enough to affect other parts of the economy, such as consumer spending, employment, inflation, and interest rates. Our project is also important for young adults who are in the market to buy an affordable home. So not only will our results help in predicting/analyzing the economy of the U.S., but they will also be a tool for young families to use. That is why our group was excited about such an idea, one that can help shape the country, but also give a realistic expectation to people looking to purchase property.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Existing Questions

There are several questions that data scientists ask about house price predictions:
1. How do different regression models compare to accuracy?
2. What features are the most predictive of house price?
3. How does model accuracy compare between different geological regions?
To answer some of them, we will be using two models to compare accuracy, and also using one variable as a control, and then removing it to see if the accuracy has changed. This should allow us to answer these questions and gain a better insight from our data.

Prior Works

There is one paper that stuck out to my group, which was published in 2023 by the Asian Journal of Research in Computer Science.[1] In summary, this paper decided to find which machine learning algorithms were the most efficient and accurate in predicting house prices, and what factors influenced prices the most. They used a public dataset from Kaggle, along with a plethora of ML models to find their results. After analyzing the data, they noticed that linear regression and lasso regression performed the best, and that the number of bathrooms and balconies showed a strong positive correlation with the price.

3 Method

Dataset

The dataset we used is the US Cities Housing Market Data by Vincent Vaseghi from Kaggle.[2]
The dataset came in the form of a tsv, which we took a sample of in the form of a csv to do our project on. There were originally 57 features in our dataset, which we made the conscious decision to reduce to 17 columns. We removed a lot of the unnecessary columns that often included meaningless or redundant data for our project.
As the 57 features provided by the database contained unrelated features like price drops (we are analyzing the final price not the amount they dropped) or redundant features such as the ones that provide month over month data. After dropping these types of unused features we ended up with the following:
median_sale_price
median_list_price
homes_sold
pending_sales
new_listings
inventory
months_of_supply
median_dom
avg_sale_to_list
sold_above_list
price_drops
city/state data
Year/Month data
Categorical variables like city, state and month were encoded using one-hot encoding (OHE) and numerical features were scaled where appropriate.

Models

We implemented three machine learning models for our house price prediction project: Linear Regression, Logistic Regression, and Random Forest. Linear Regression served as our baseline model, offering a simple and interpretable way to estimate housing prices by modeling a linear relationship between input features and the target variable. Although it assumes linearity, it is useful for quick insights and comparison with more complex models. Logistic Regression was used to model a binary classification problem by defining a new target variable, whether a house sold above its list price. This model provides probabilistic outputs and is particularly effective when the business goal involves price tier prediction rather than exact pricing. Our most accurate model was Random Forest, an ensemble method that builds multiple decision trees using random subsets of data and features, and averages their predictions. It effectively captures non-linear relationships and interactions between variables, handles missing data and outliers, and provides feature importance scores. Random Forest's robustness and high predictive performance made it the most beneficial model in our study.

Evaluation

We evaluated each model using an 80/20 train-test split, ensuring that the testing set remained unseen during training to assess generalization performance. For the two regression models we measured performance using R² score and Mean Squared Error (MSE). R² indicates how much variance in the target variable is explained by the model, while MSE quantifies the average squared difference between predicted and actual prices.

4 Results

Our model performance varied across methods, reflecting the nature of the data and the suitability of each algorithm. With all features included, the linear model achieved an R^2 score of 0.534 and an MSE of 2.52×10^{10}. This insight led us to adopt Random Forest Regression in search of possible better results with an R^2 score of 0.663 and an MSE of 1.82×10^{10}. Due to its ability to handle nonlinear relationships and isolate useful splits through decision trees, Random Forest was able to manage the complexity of the one-hot encoded city features effectively, outperforming the linear model.

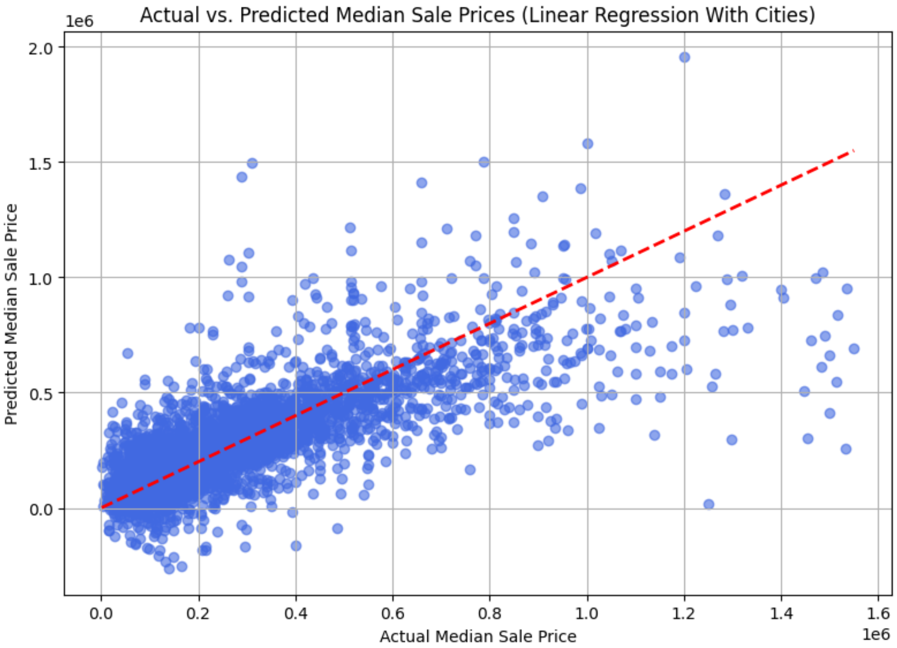
To further investigate the effect of high-cardinality geographic data, we trained both models again without city features, removing both the raw city column and its one-hot encodings. For Linear Regression, performance improved slightly: R^2 increased from 0.534 to 0.647, and MSE dropped from 2.52×10^{10} to 1.91×10^{10}, suggesting that the exclusion of city data reduced overfitting and improved generalization. However, Random Forest performed about the same without city data, with R^2 moving to 0.666 and MSE decreasing slightly to 1.8×10^{10}. This result highlights Random Forest's reliance on geographic features to partition the data effectively and its strength in handling high-cardinality, sparse features. The contrast between these outcomes illustrates the importance of feature selection and model compatibility: while Linear Regression benefits from dimensionality reduction, Random Forest leverages complex, high-dimensional inputs to capture nuanced patterns in the data.

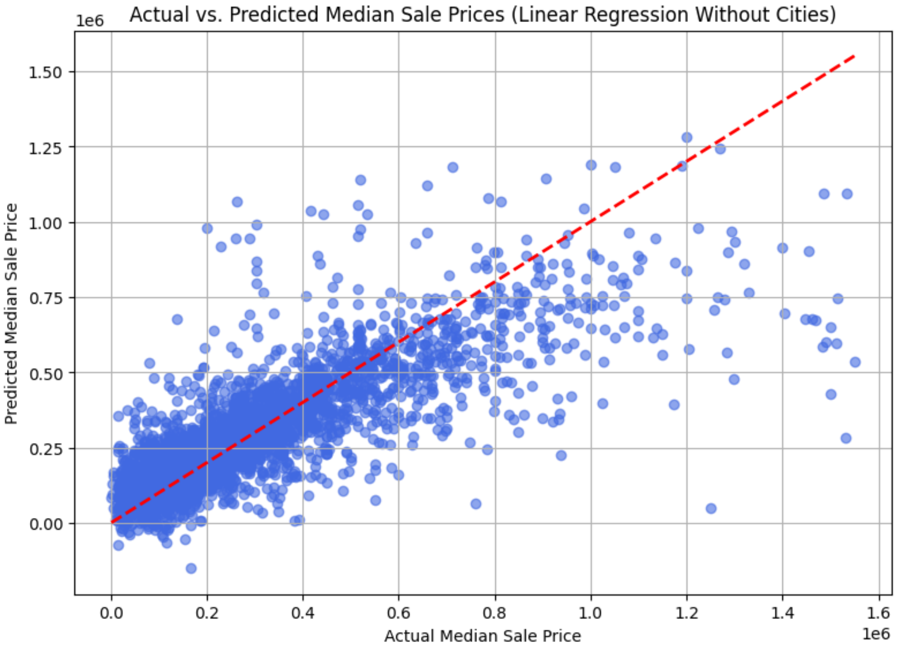
For the classification task, we used Logistic Regression to predict whether a home sold above its list price, framing it as a binary classification problem. The model's overall performance was modest, achieving an accuracy of 0.566 both with and without city features. The precision, recall, and F1-scores showed an imbalanced result: for class 0 (not sold above list), the model achieved a precision of 0.61, recall of 0.78, and F1-score of 0.69, while for class 1 (sold above list), the precision dropped to 0.40, recall to 0.23, and F1-score to 0.29. This indicates that the model was biased toward the majority class and struggled to correctly identify homes that sold above list price. Removing city features had no measurable effect on classification performance, suggesting that location data either lacked predictive power for this binary outcome or was already underutilized by the model due to convergence limitations.

Additionally, the model issued a convergence warning, indicating that it failed to fully optimize within the default number of iterations. This suggests the need for further preprocessing, such as feature scaling or solver adjustment, to improve optimization and performance. Despite its interpretability and fast runtime, Logistic Regression underperformed on this task relative to our regression models and highlights the challenges of applying linear classifiers to imbalanced, complex real estate data. In summary, Random Forest proved to be the most effective model for regression due to its ability to handle nonlinearity and high-cardinality features. While Logistic Regression provided interpretable outputs for classification, its performance was limited by class imbalance and convergence issues. We used cross-validation throughout to ensure model stability and prevent overfitting. These results highlight the importance of aligning model complexity with data structure when working with real estate prediction tasks.

5 Visualization

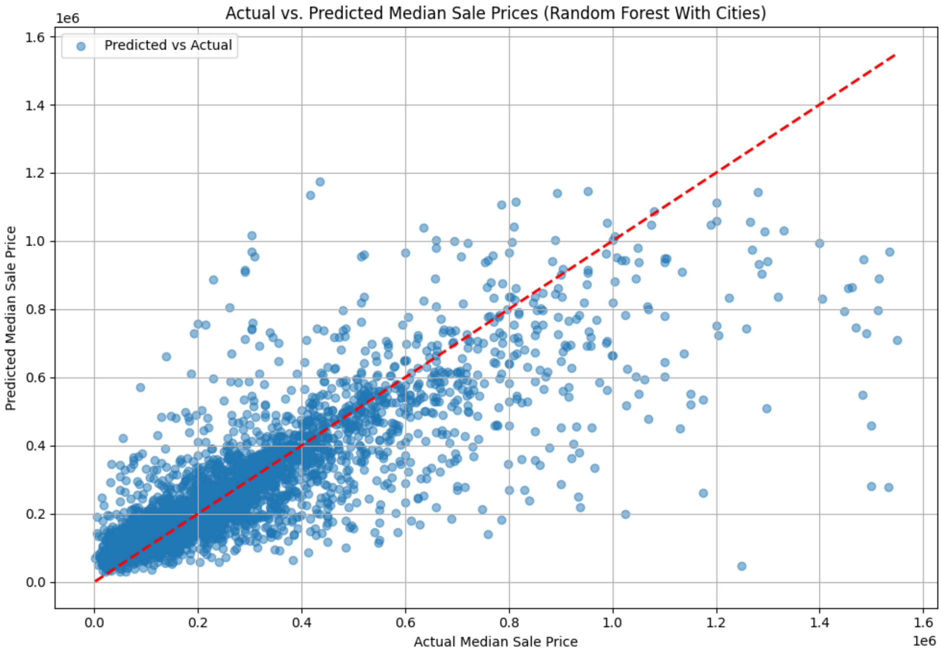
Linear Regression

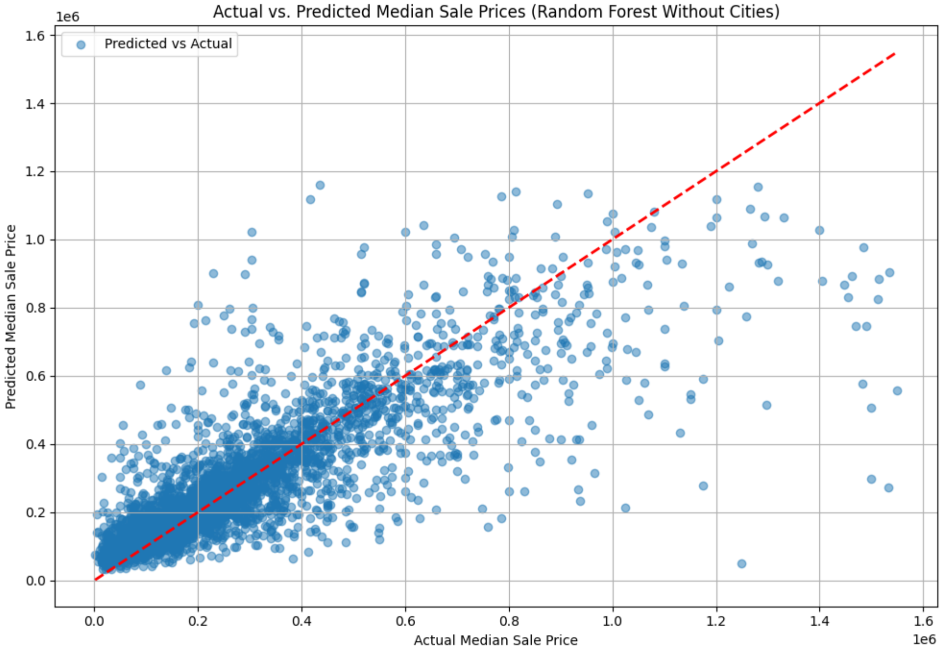




For Linear Regression, the model that used without cities performed visibly better compared to the model that used cities data. The blue dots estimate how closed the prediction was to actual. Based on the graphs, our model understands when to price homes higher however, it seem to under predict a lot of the home prices. Because home prices are not linear the model may be underfitting because the data itself is not matching up. Our model seems to be able to predict average price houses however, there seem to be more difficulty when it comes to predicting more expensive properties.

Random Forest Regressor





If the dot is below the red line that means the model under-predicted, if above the model over-predicted. For Random Forest Regressor, we actually encountered a lot of the same problems where the more expensive the property is the more unlikely our model is able to accurately predict.

Analysis

The performance differences across models reflect the strengths and weaknesses of each model. Linear Regression struggled with high-dimensional data due to one-hot encoding of city variables, but performed well when the city features were removed. This reinforces the need for simple models to have clean, low-noise input spaces for linear regression models.

Random Forest, by contrast, performed best with city data included, highlighting its ability to capture localized, nonlinear effects and partition complex feature interactions. However, when location data was excluded, performance dropped sharply—revealing the model's dependence on geographic signals in housing prices.

Across all models, we observed systematic underprediction of expensive homes. We believe that the reason why we have trouble predicting expensive houses may be due to the metro region where houses are located because a house in New Jersey who is in the same metro region as Manhattan may not be as expensive as housing in Manhattan. Other factors may include a lack of data for more expensive houses as we had to cut our dataset down to 20000 entries in order to have a viable training time. All in all, our model seems better for prediction mid to low price range houses.

Logistic Regression also suffered from class imbalance and convergence issues, limiting its ability to generalize. Its inability to improve with city data suggests that whether a house was sold above list or not may depend more on unobserved behavioral or timing factors than geographic ones.

6 Discussions

We expected our results to be at least 50% accurate in comparison to the dataset. As we got further along into the project, we realized the numerous scales and variables that play into the housing market. We realized that due to our machine power, it wouldn't be possible to create a complex model, and we had to work with what we were given. In our models, we never seem to break into the 70% R^2 value. However, for our actual results, we realized that the outcome of the predictor models is highly dependent on the features we input. For example, the random forest regressor worked better with the addition of city information compared to linear regression, where city information gave us a worse R^2 score. We debated a lot about whether to include the city as a feature for our model due to the vast number of them that resulted from one-hot encoding, but realized that they had an impact on our models. Our original project proposal included logistic regression however, we quickly realized that logistic regression is a classification model that doesn't help in our predictive question. However, we incorporated the logistic regression model into our project, so it answers the question of whether a house will sell above the listing price. Although this model doesn't answer the question of the price of a house, it still pertains to the price of a house by assessing the chances of a house being sold for much more than originally intended. Overall, this project introduced a lot of difficulties for us because we had a hard time finding models that worked with our data, as housing prices are extremely volatile.

7 References

1. (PDF) Machine Learning Approach for House Price Prediction, www.researchgate.net/publication/371602053_Machine_Learning_Approach_for_House_Price_Prediction. Accessed 9 May 2025.
2. Vaseghi, Vincent. "US Cities Housing Market Data - Live Dataset." Kaggle, 8 Apr. 2025, www.kaggle.com/datasets/vincentvaseghi/us-cities-housing-market-data/data.