Rodolfo Rivera, Daniel Gilberto
Mansoureh Lord
Comp 541
1 May 2024

Predicting Home Prices In Los Angeles County

INTRODUCTION

Los Angeles home prices have been on a steady incline since the late 90's when viewed at a glance, to the point which peaks interest in research and the prediction of the future housing climate. As the housing market grows with the generations, younger buyers typically expect to afford housing in the future. However through research and data mining, we will discover the trends which may prove or disprove this idea. The determining factors in predicting a house price varies from the area a home may be in, the amount of bedrooms a home has, the amount of bathrooms a home has and how physically big the home is. Accurate prediction of home pricing is very valuable information as it can help someone make a lot of money or save someone a lot of money. This project presents a comprehensive study in gathering, cleaning and using machine learning models to predict house prices with a focus in the county of Los Angeles using a myriad of common and specific techniques which relate to data mining.

The primary objective of this project is to predict the house listing prices in Los Angeles county using data acquired from the website crawlers, cleaning, preprocessing, visualizing and using machine learning models. We will employ models such as linear regression and gradient boosting to find the trends in listing prices. These models will provide useful insight for people who are in the market for a new home or for people who are in real estate to find favorable homes, tailoring this data to the greater Los Angeles county area.

We acquired our data from two separate websites in order to provide insight from two independent sources. This is to prevent skewed or biased data which may be incurred from using a singular knowledge base. The two websites we set our sights on were redfin.com and compass.com using website crawlers written in R, with one crawler written by each. The data gathered from the two crawlers were formed into separate data frames, listed with its address, zip code, number of bedrooms, number of bathrooms, area in square footage, and the price it sold. We gathered the data of sold homes from the past two years to mitigate the disparity of home prices throughout many years, since home prices have been in a steep incline since 2020 and would have a big effect on the accuracy of the machine learning models. We then cleaned the data and prepared it to merge together.

Initially, one dataset had gathered over 62,000 rows and the other dataset gathered over 24,000 rows of data. After the data preprocessing, our data set consisted of over 82,000 total rows and 6 columns. The 62,000 rows of data were gathered from the listings on redfin.com and the 24,000 rows of data were gathered from the listings on compass.com. Both of these listings contained housing data from Los Angeles county as a whole. The data frame was created from the data scraped on these websites in February, 2024.

The process of gathering this data was challenging as there were many restrictions imposed by every real estate website that provided listings of homes in the United States. These websites limited or blocked requests from automated machines like crawlers to prevent people from mass gathering data as it is very valuable information that they would rather charge people to gain than get for free. Redfin.com for example, you're able to download loads of data but only if you are a paying member of their website. With these barriers we had to scrape the data from these websites in areas where our requests were not being blocked. This included the initial search page in which all the listings were posted on the page, with their limited information showing. Using this method, we were limited in the amount of information we were able to extract from each home. The available information included the address, the area in square footage of the home, the number of bedrooms it had, the number of bathrooms it had, its zip code, and its listed price.

Table 1: A subset of the homes dataset post-processing

address	zip_code	num_bed	num_bath	home_area	price
2602 Ladoga Ave Long Beach	90815	3	2	1394	987000
25909 Burke Pl Stevenson Ranch	91381	4	3	3058	1130000
220 S Berkeley Ave Pasadena	91107	2	2	956	820000
7838 Ellenbogen St Sunland	91040	3	2	13994	945000

In the process of data cleaning, we encountered many issues with our data set that needed to be dealt with. These issues included; missing values, numeric values that were strings, mixed measurements (sq ft and acres) and outliers. For our missing values in the bedroom and bathroom column, we decided to use the median of the respective columns to fill in the missing values due to median being robust to data that is skewed. Since we had a lot of data that were outliers, with homes in the higher end having over 15 bedrooms and bathrooms. In regards to data missing home area values, we had a total of 142 instances and decided to use a linear regression model to fill in these values. Finally the data with 3 or more missing values were completely dropped from the data set to reduce inaccurate real world data, to preserve the integrity of the data set and to prevent any false patterns that may arise. Our data set had a total of 74 instances with 3 or more missing values that were dropped from the set.

When merging the two data sets, we standardized the data into a consistent format. This included we had the same variable type, between numeric, character, and integer to ensure there would be no issues when combining the data. After combining the two data sets, we then removed the duplicates which brought our total data set to 82,000 instances before the removal of outliers. Finally we were left with the task of removing outliers to improve the accuracy of our machine learning models and increase its robustness. For this task we decided to use Isolation Forest, which at its core, uses binary search trees to isolate data and efficiently identify outliers. In a study done in 2023 by Liu, Tao, et al. titled “Layered Isolation Forest: A Multi-Level Subspace Algorithm for Improving Isolation Forest.”. In this paper, the authors go in depth on how isolation forests work and describe the pros and cons of using this technique.

The advantages of using isolation forest to identify outliers include, efficiency and effectiveness in regard to large-scale data, due to randomly selecting partitions and constructing binary trees that effectively identify anomalies in the data set. Due to its exceptional performance and linear time complexity, isolation forest is an algorithm that is respected and attractive for researchers.

The disadvantages of using isolation forest include misjudgement of some data points due to unused information that leads to “ghost” phenomenon, which is missed outliers. This labels some of the data points unfairly and inaccurately. This leads to an anomaly score that does not match the distribution of data. The algorithm shows low interest in outliers that are close to normal data points, which results in a decreased accuracy of identifying local anomalies. Lastly, the model uses random values for partitioning, which results in missing important information for detecting outliers.

To visualize how isolation forest affected our data, a price to home area chart was created before and after using isolation forest and shown in figure 1.

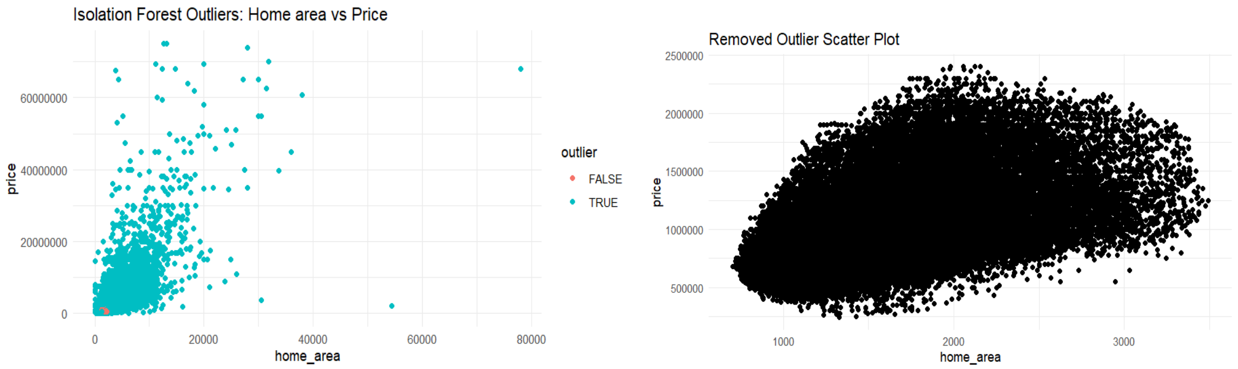


Figure 1: Scatter plots showing the before and after isolation forest on our data set.

In Figure 1, we see how many data points were scattered and were outliers to our data set. The data range for price was reduced from 80,000,000 to 2,500,000. The home area range was reduced from 80,000 to about 3500. It shows how dense the data is in such a short range, letting a machine learning model become more accurate than with data with outliers. Isolation forest identified about 28,000 outliers, which we then removed to get our final clean data set with 53,401 instances.

In order to better understand our data, we turned to R to generate visualizations of certain categories. Below are the visualizations of the data before applying any data mining techniques along with a brief description of each.

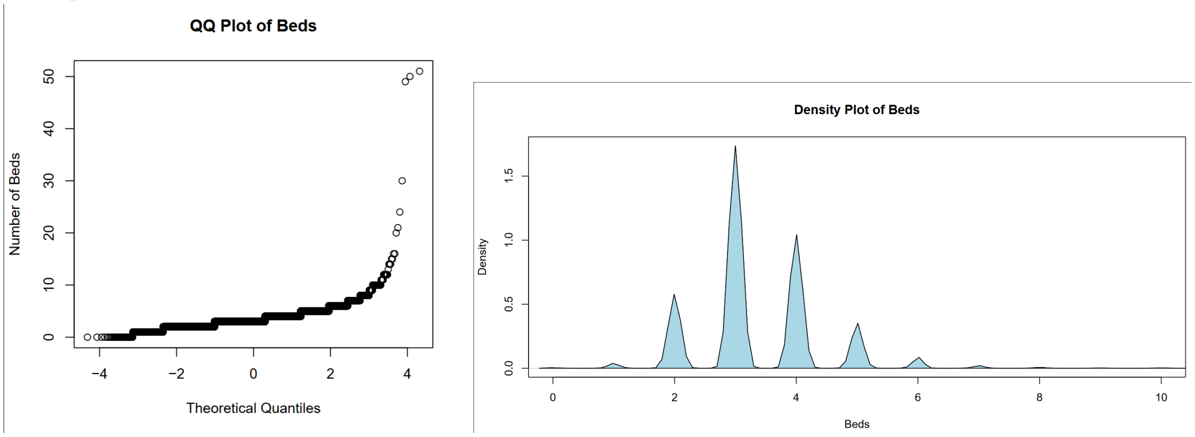


Figure 2a: Data Visualization, Number of Beds in data set

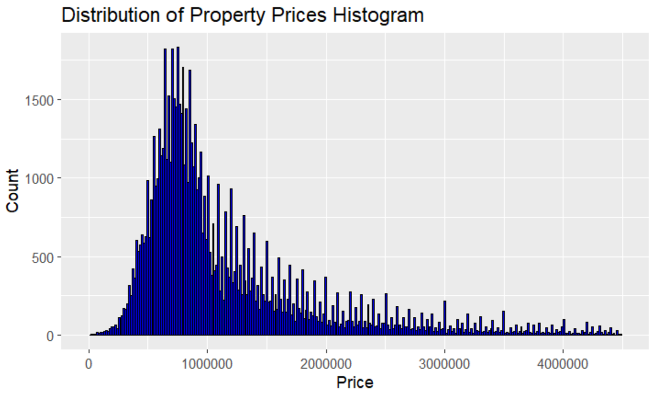


Figure 2b: Data Visualization, Count of Properties vs Property Prices Prices

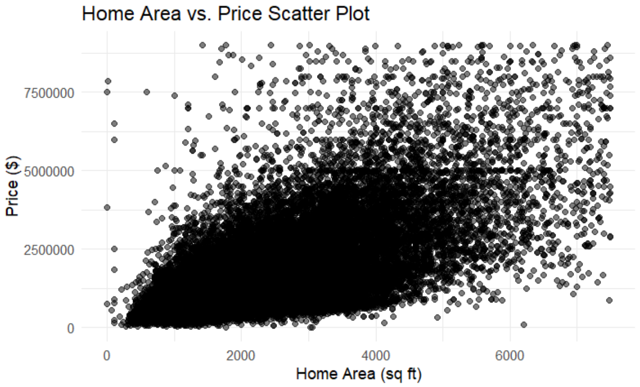


Figure 2c: Data Visualization, Home Area vs. Square Footage Scatter Plot

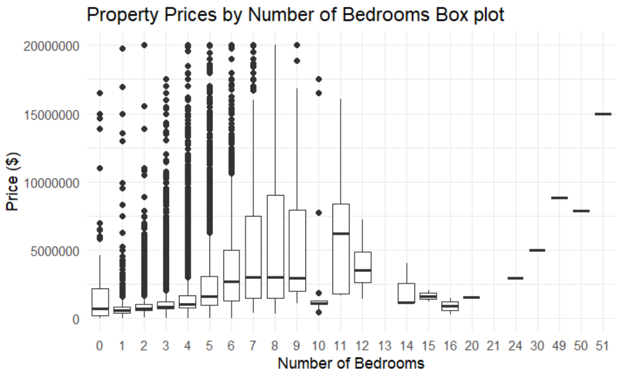


Figure 2d: Number of beds vs. Property price

Looking at these visualizations at a glance show that a majority of homes are within the 500,000 - 1,000,000 range. Another interesting trend is that the more bedrooms there are the pricer the house is. A surprising feature of the visualizations show that the square footage does not seem to have a tremendous impact on the price of the home. This can most likely be a correlation of the ZIP code of the property. A 1000 square foot home in Beverly hills can be worth a lot more than a 1,000 square foot home in palmdale.

RELATED WORK

Work 1: Predicting listing prices in dynamic short term rental markets using machine learning models
https://arxiv.org/pdf/2308.06929

Work 2: Prediction of real estate prices with data mining algorithms
https://www.researchgate.net/publication/342782159_Prediction_of_real_estate_prices_with_data_mining_algorithms

Work 1 served as an inspiration to the structure of this report. It goes on a very in-depth analysis of a similar topic related to the prediction of listing prices in a market. The work specializes in short term rental markets like Airbnb, whereas our project is in regards to the prices of homes in LA County. They also use similar techniques in which we settled on to provide low-error predictions when utilizing the models.

Work 2 is a report on real estate price prediction and which data mining techniques would best be suited. The report covers three different data mining techniques: random forest, linear regression, and gradient boosting and then reports back with the positives and negatives of each, as well as which techniques were the most well-suited for the topic at hand.

METHODS

Considering the data we had and the objective of our project, we decided to use Linear Regression and Extreme Gradient Boosting as our machine learning algorithms. Linear regression models are often employed in price prediction tasks and serve as a good model to use for our project. It performs well for linearly separable relationships between dependent and independent variables. With the advantages of using linear regression, it also has its disadvantages. Since linear regression uses a linear relationship between the variables and target, it may not fully capture the complexity of how all the variables contribute to predicting home prices. It is sensitive to outliers because they distort the model performance. It also is susceptible to noise and overfitting, which we will combat with a cross validation method.

Using linear regression, we were able to achieve accurate results by being able to include multiple relevant parameters that would contribute to the prediction of home prices. For the linear regression model, we used a 80/20 split for training and testing. This means that 80 percent of our data base was used to train the model and 20 percent of the data was used for testing purposes. The general equation for linear regression with multiple variables is

$$Y = \mathrm{b}_0 + \mathrm{b}_1\mathrm{x}_1 + \mathrm{b}_2\mathrm{x}_2 + \ldots + \mathrm{b}_n\mathrm{x}_n$$

Extreme gradient boosting is a data mining technique which is primarily a linear regression type technique, but addresses slight categorization as well. This technique is known as an ensemble method because of its use of multiple weak learners. It works through the use of these learners to solve many small problems and utilize the results for the next iteration of mining. These weak learners are simple decision trees and are meant to perform both quickly and efficiently with large data sets. Once the model is complete, it provides valuable insight into how the features of a house will be categorized. In order to use this technique in our project, we had to designate categories which make sense to the results we are attempting to discover. The categories we used are as follows:low-end, affordable, high-end, and luxury. These price ranges are defined as “0-99,999” for low-income, “100,000 - 499,999” for affordable, “500,000 - 999,999” for high-end, and greater than 1,000,000 will be luxury.

Just as in linear regression, this model also uses the concept of data splitting to allocate a certain percentage of the data set to training and the other part of the data set for testing the trained set against a control. For this technique, we used an 80/20 split like in the linear regression portion of this project to keep uniformity of technique parameters in order to compare.

In order to evaluate the performance of our linear regression model, we will use k-fold cross validation. This method provides a more reliable estimate of the model's performance compared to a single 80/20 split of train and testing data. The research done by Xiong, Zheng, et al. (2020) in the article “Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery Using K-Fold Forward Cross-Validation.” explains how k-fold cross validation works. It splits the data into k subsets or folds and is usually selected randomly. The model then trains k times, each time using 1 less subset/fold and the remaining of the data as validation data. Using this method, it provides a more robust evaluation of the models general performance.

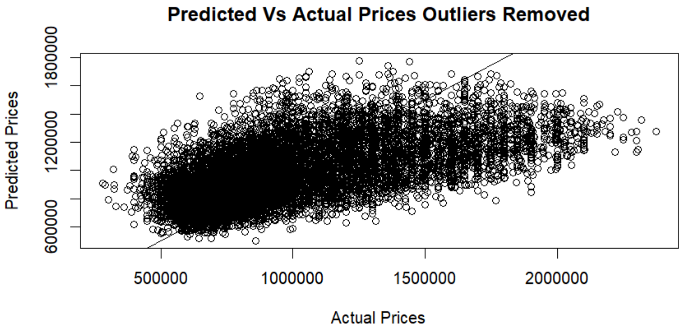
EVALUATION

Our linear regression model had a good prediction performance, predicting on average about 7.56% from the actual home listed price. This gave us a residual standard error of 288,900, which represents the typical difference between the actual home price and the predicted home values. The model gave us a multiple R-squared value of .3481, which indicates that there wasn’t significant variability in the home pricing. The higher the multiple R-squared is, the more there is fluctuation in prediction outcomes. This could be due to many real world factors that affect home prices, and a very competitive housing market that would pressure real estate agents to competitively price their home. The adjusted R-squared had a value of .348, which indicates that there were no negative effects in using the parameters used.This tells us that the model was using all the parameters given, and using them had a significant impact on the prediction of a home. This means that our linear regression model would likely benefit from being given more parameters to judge the price of a home to make it more accurate. Finally we get a p-value of < 0.00000000000000022, which tells us that the parameters definitely had a significant effect on determining the pricing of a home. The lower the p-value, the lesser probability of the null hypothesis, and indicates that any observed difference in predictions is not due to chance.

Table 2: Shows the results of the linear regression model performance

	Linear Regression	Extreme gradient boosting
RSE	288,900	312,600
Multiple R-squared	0.3481	0.3963
Adjusted R-squared	0.348	0.396

Figure 3: Visual representation of the linear regression model



The k-fold cross validation had similar results to the 80/20 split training. With k-fold cross validation, we received a Root Mean Squared Error of 288,652, an R-squared of .3485 and a Mean Absolute Error of 221,390. The Root Mean Squared Error is the square root of the average squared error between the predicted price and actual price. It is more sensitive to large errors because of its squaring operation. The Mean Absolute Error is the average difference of price between the actual and predicted price. The lower Mean Absolute Error of 221,390 compared to the Residual Standard Error of the linear regression model of 288,900 indicates that the linear regression model has generalization performance, it would perform well on unseen data and suggests that it is not overfitted.

Table 3: Shows the results of k-fold cross validation on linear regression model

	K-fold Cross Validation
RMSE (Root Mean Squared Error)	288,652
R-squared	0.3485
MAE (Mean Absolute Error)	221,390

CONCLUSION

This project has been very insightful in two main aspects: one being that actual topic we were researching, and the other being all that we have learned while accomplishing our goals. Through data preparation, data cleaning, and integrating regressive data mining techniques, we were able to produce a solid foundation of predicting housing prices given the recent and current market values of the area.
We were able to utilize and fine tune both linear regression and gradient boosting as our two data mining techniques to accomplish the task of predicting L.A. county real estate prices.
Given the results of both techniques, we are able to come to the consensus that homes in Los Angeles county are within the high-end range and appear to be rising. The biggest contributing factor to this inflation is the number of beds and baths. These two shows trends of having pricier tags.
For the future of this project, it would most likely be best to use a neural network regression model for predicting L.A. County Housing prices. Neural networks have demonstrated high proficiency in capturing intricate patterns and relationships in data. It would also be best to gather more data for parameters to further improve the accuracy of the model and give it more data to work with.

Works Cited

Liu, Tao, et al. “Layered Isolation Forest: A Multi-Level Subspace Algorithm for Improving Isolation Forest.” *Neurocomputing (Amsterdam)*, vol. 581, 2024, https://doi.org/10.1016/j.neucom.2024.127525.

Xiong, Zheng, et al. “Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery Using K-Fold Forward Cross-Validation.” *Computational Materials Science*, vol. 171, 2020, pp. 109203-, https://doi.org/10.1016/j.commatsci.2019.109203.

Chapman, Sam, Seifey Mohammad, and Kimberly Villegas. "Predicting Listing Prices In Dynamic Short Term Rental Markets Using Machine Learning Models." arXiv preprint arXiv:2308.06929 (2023),https://arxiv.org/pdf/2308.06929

Uzut, Gulsum & Buyrukoglu, Selim. (2020). Prediction of real estate prices with data mining algorithms. Euroasia Journal of Mathematics Engineering Natural and Medical Sciences. 7. 77-84.

Khare, Sandali, et al. “Real estate cost estimation through data mining techniques.” *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, 1 Mar.2021, p. 012053, https://doi.org/10.1088/1757-899x/1099/1/012053.