Predictive Models on Housing Prices (Using Zillow Housing Data)

Data Bootcamp Final Project

Bipana Bastola, bb3489

Introduction/ Problem Statement

The real estate market is a significant constituent of the economy and individual financial choices/decisions. With high fluctuations in the housing prices across regions, predicting the home prices is an area of great importance. Accurate predictions serve as valuable tools for real state agents in addition to easing the decision making processes for both the buyers and the sellers. This project aims to predict the housing prices in the United States using various machine learning models using the data from Zillow Listings. The dataset contains information about the properties across the US states including attributes such as the number of bedrooms, bathrooms, the year of construction, lot size, its price, and the qualitative descriptions of the properties.

The analysis in this project encompasses the following key steps:

- Data cleaning and preprocessing
- Exploratory data analysis
- Implementation of multiple predictive models
- Evaluation of the model performance
- Feature importance analysis
- Incorporation of sentiment analysis

The primary objective of the project is the develop as accurate predictive models for housing prices and determine the most influential factors affecting those prices. Models such as Linear Regression, K-Nearest Neighbors (KNN), Decision Trees, and Random Forest were developed and assessed for their ability to accurately predict property values. Additionally, sentiment analysis was conducted on property descriptions to examine whether linguistic nuances influence housing prices. By combining statistical analysis and machine learning techniques, this study aims to provide insights into the housing market and propose a framework for future applications in real estate analytics.

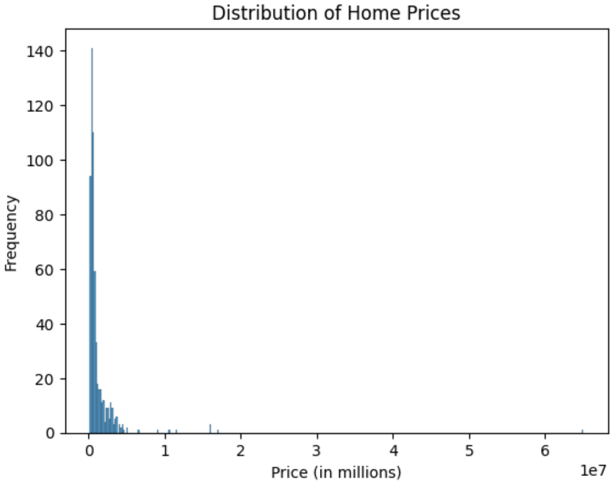
Dataset Description

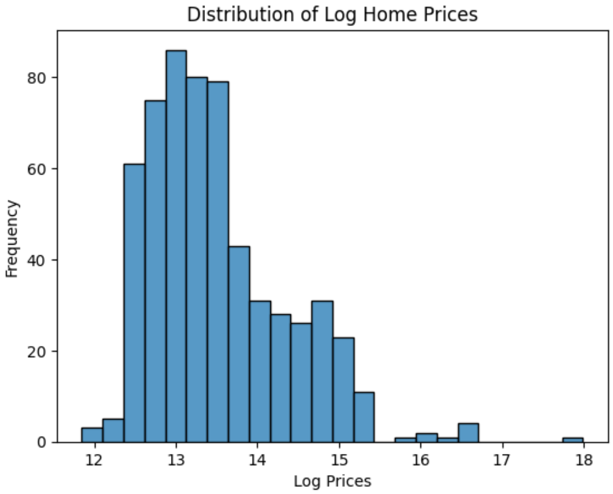
The data for this project is the Zillow housing listings data and was accessed from Bright Data [1]. The dataset provides the a comprehensive view of the properties built after 2015 that are available for sale. The downloaded dataset contained 1000 entries in the initial form. It included a mix of numerical, categorical and textual data. The key attributes included city, state, home status, address, and the property details such as bedrooms, bathrooms, living area and prices. It also provided insights into tax assessments, descriptions, etc.

Data Cleaning and Preparation

Upon exploration, the dataset contained missing values and irrelevant features. Attributes such as the ‘zpid’, ‘Address’ and other irrelevant descriptive fields were excluded from the analysis. Properties with missing values in critical columns like price, bedrooms, and bathrooms were removed. Furthermore, properties with a price of zero were filtered out, as they would skew the predictive models. After this cleaning process, the dataset was reduced to 591 entries, retaining only the most relevant rows and columns.

I decided to drop the missing values rather than imputing them to maintain data integrity. Given the small size of the datast post-cleaning, retaining only complete and meaningful entries was important to ensure reliable model training and evaluation.



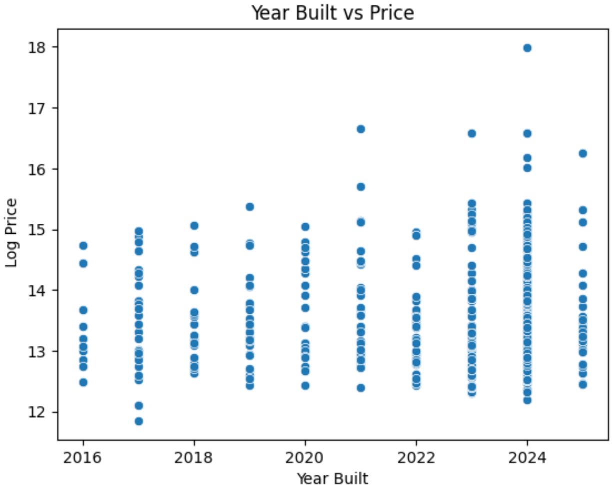


Since property prices exhibits a right-skewed distribution, a log transformation is applied to normalize the data. This transformation ensures that models relying on assumptions of normality, such as Linear Regression, could perform more effectively. The cleaned dataset provids a robust foundation for further analysis and modeling.

Exploratory Data Analysis

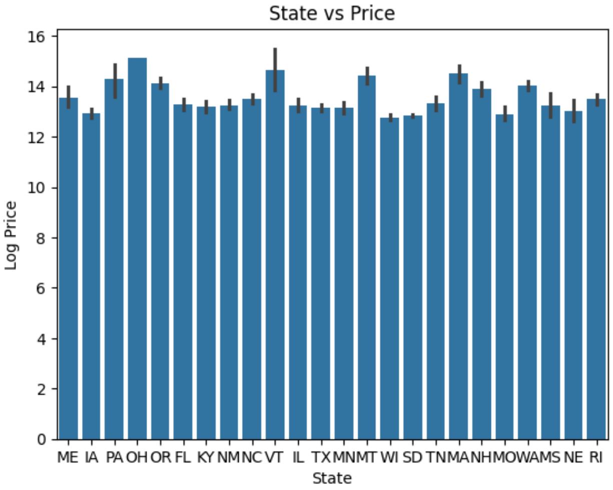
Exploratory Data Analysis (EDA) is an important step in understanding the data superficially. Several visualizations are created below to understand the data distribution and the relationships.

1. Scatter plot of year built vs. price



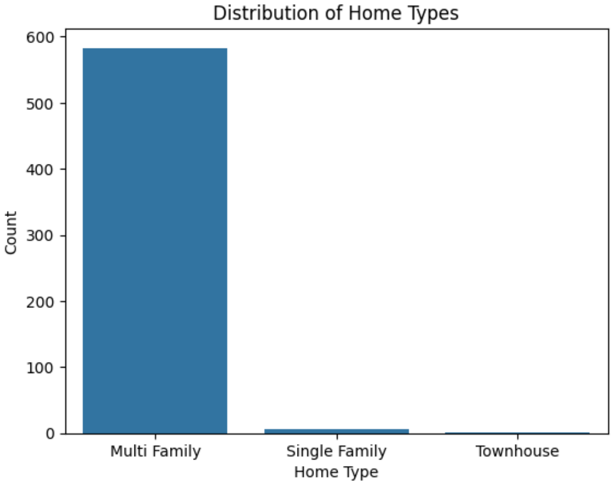
This scatter plot of year built against the log-transformed price reveals a weak positive trend, suggesting that the newer properties tend to have higher prices. This trend aligns with expectations, as newer homes often require less maintenance and incorporate modern features that appeal to buyers.

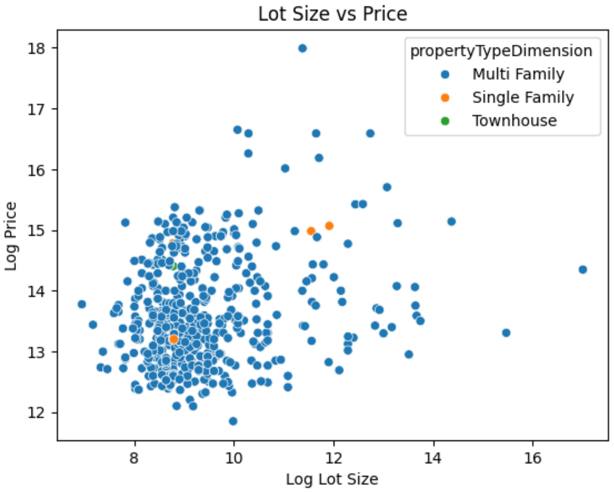
2. Bar plot of average price by state



This analysis of property prices by state highlights substantial regional variances. States like Massachusetts (MA), Vermont (VT) and Ohio (OH) exhibit higher median prices compared to others. This reflects the influence of regional economic conditions, demand, and local amenities on housing prices. For example, states with thriving job markets or desirable living conditions often command higher property prices.

3. Distribution of home types





The distribution of the property type suggests that in the dataset the majority (583 out of 591) consisted of multi-family homes, with single-family homes and townhouses making up a small proportion. As we see in the scatter plot, single houses and townhouses do not seem to have distinct patterns. Yet, we can not say for certainty that the type of homes do not have any affect on the prices. The sample in this case is too little for us to draw any definite conclusions about relationship between the type of home and its price.

Baseline Model: Mean Price prediction

To establish the baseline criterion for evaluating the predictive modles, a baseline model using te mean log price is implemented. This model assumes that all the properties will be priced at the average log price (i.e. 13.513). This results in the **Mean Squared Error (MSE) of 0.721**.

Modeling and Interpretations

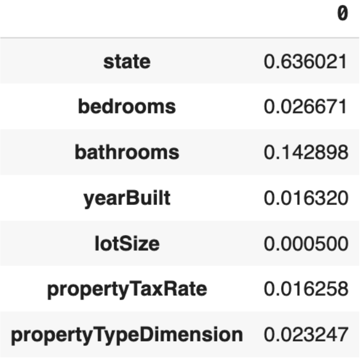
I implemented the following models and the models are evaluated using Mean Squared Error (MSE) on both training and test sets. A train-test split of 80-20 was used for all the models.

Pre-processing

Before fitting the predictive models, categorical variables (‘state’ and ‘property type’) are encoded suing one-hot encoding to convert them into numerical representations. The predictors are ‘state’, ‘bedrooms’, ‘bathrooms’, ‘yearBuilt’, ‘lotSize’, ‘propertyTaxRate’, ‘propertyTypeDimension’ while the response variable is ‘LogPrice’. The preprosessing ensures that the data is ready for the predictive models.

1. Multiple-Linear Regression

The the Mean Squared Error for the train set (0.3931) and test set (0.3024) both outperformed the baseline mean prediction model. However, its relatively higher error suggested that a linear approach might not capture the complex relationships within the data.

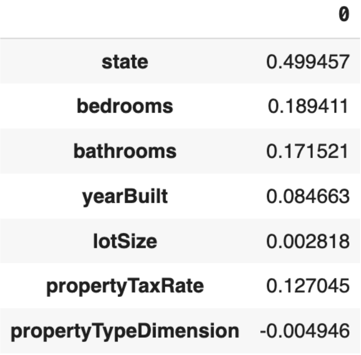


	0
state	0.636021
bedrooms	0.026671
bathrooms	0.142898
yearBuilt	0.016320
lotSize	0.000500
propertyTaxRate	0.016258
propertyTypeDimension	0.023247

From the regression model, we see that the geographical location (i.e. State) is the most important factor in determining the housing prices followed by the number of bathrooms in the house. The least important feature, as per the model, is the lot size.

2. K-Nearest Neighbors (KNN)

KNN regression is used next to capture non-linear relationships. Hyperparameters are optimized through GridSearchCV, and it is found that the model with 5 neighbors is the optimal choice. The model suggests a MSE of 0.2476 for the train set and a MSE of 0.2086 for the test set both of which outperform the Linear Regression model.

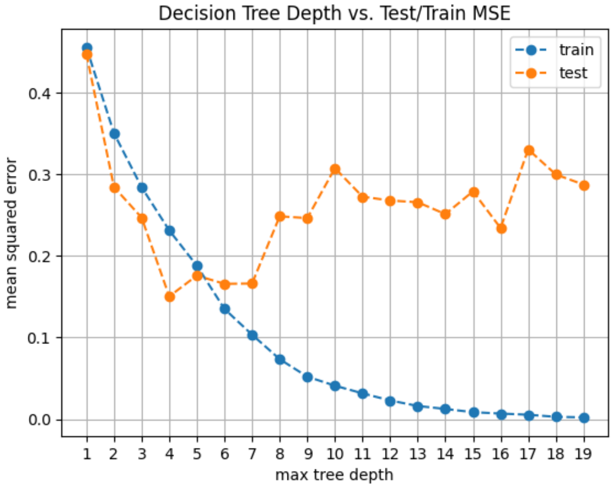


	0
state	0.499457
bedrooms	0.189411
bathrooms	0.171521
yearBuilt	0.084663
lotSize	0.002818
propertyTaxRate	0.127045
propertyTypeDimension	-0.004946

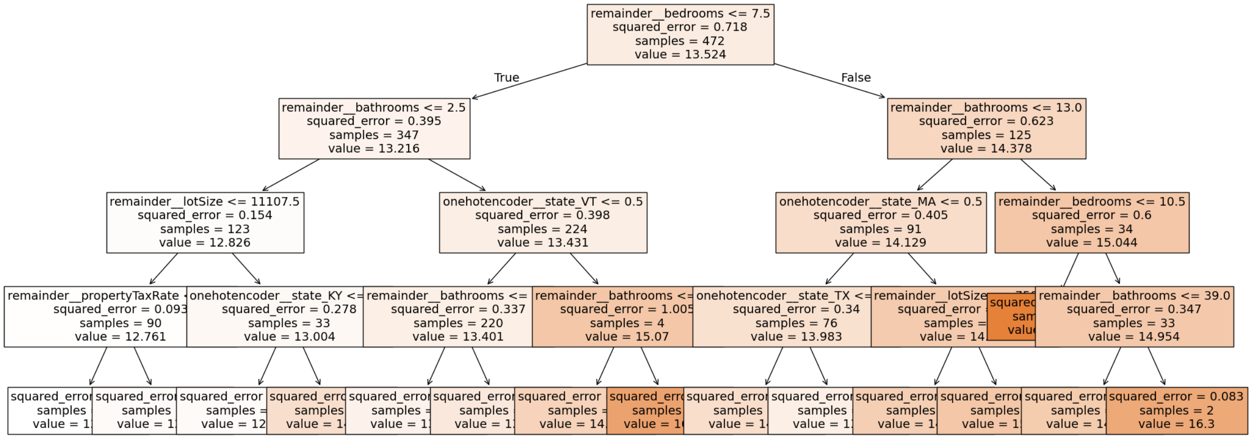
Again, the State where the property lies is the most important feature, followed by the number of bedrooms and bathrooms. Property Tax Rate also emerges to have a strong influence on housing prices.

3. Decision Tree Regressor

The decision tree model is built to model the non-linear interactions and feature hierarchies following the KNN model.



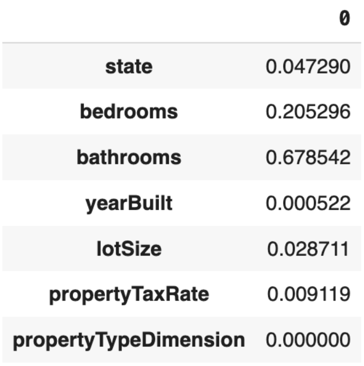
The optimal height of the tree is determined to be 4, with a good balance omitting both under fitting and over fitting. Decision tree regression with height 4 performed than all the previously fitted linear regression and KNN models surpassing MSE for both the train sets. The MSE for the train set is 0.2314 while it is 0.1424 for the test set.



This representation of the tree shows the importance of features like the number of bathrooms and bedrooms in determining the house prices. The hierarchical structure of Decision Trees is effective for capturing interactions between categorical and numerical variables.

4. Random Forest Regressor

The Random Forest model is buit to enhance predictive performance through aggregation. Using GridSearchCV, the model's hyperparameters, the number of estimators and the max depth, are tuned. The model configured that the optimal choices are 50 estimators with the depth of 3. The MSE for the train set is 0.2352 while it is 0.1766 for the test set. The random forest model, with optimal parameters of 50 estimators and a max depth of 3, performs well but does not surpass the decision tree model. However, the random forest model demonstrates the robustness to overfitting.



	0
state	0.047290
bedrooms	0.205296
bathrooms	0.678542
yearBuilt	0.000522
lotSize	0.028711
propertyTaxRate	0.009119
propertyTypeDimension	0.000000

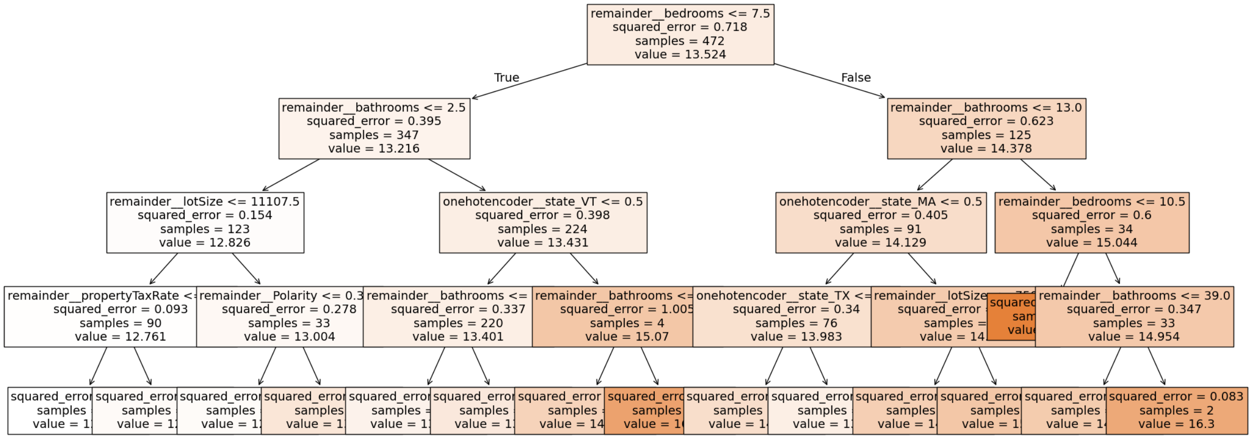
It gives the results that are slightly consistent with the other fitted models. This model suggests that geographical location has little effect on the house prices and the most influential factors are the number of bathrooms and bedrooms followed by the property tax rate.

Sentiment Analysis

In addition to the predictive models, I wanted to check if textual description of the houses had any impact on the prices. Hence, I analyze the description for sentiment using the TextBlob Library [2]. The polarity score [-1, 1] is used which quantifies the emotional tone of the descriptions. -1 indicates a very negative tone while 1 indicates a very positive tone.

Firstly, I add the polarity score to the dataframe and use it in the best performing model, i.e. Decision Tree Regressor with height 4, to see if sentiment has any effect on prices.

The MSE for the train set with polarity of the sentiment included is 0.2304 and that for the test set is 0.1531. While the MSE for the train set is slightly smaller than the MSE for the train set from the decision tree without sentiment, the MSE for the test set is higher. The higher MSE for test set shows that sentiment polarity score does not significantly improve the model's predictive performance. This suggests that the inclusion of sentiment scores may have introduced noise into the model rather than providing meaningful predictive value.



This tree representation further supports this finding, as polarity scores rankes among the least significant predictors compared to attributes like the number of bathrooms, bedrooms, and location.

Results and Interpretations

The performances of the models varied. All of the models I fitted performed better than the baseline mean price prediction model and the Decision Tree model with a depth of 4 was the best-performing model for predicting housing prices. Its ability to capture the hierarchical interactions between the features is probably what makes it the best performing model.

Location (State) is consistently the most important factor in determining housing prices across all models followed by the number of bathrooms and bedrooms. Property tax plays a significant role and the year of construction and lot size relatively are least impacting predictors of the house prices. Property type had the minimal effect on prices across all models. However, this could be because of the highly skewed frequency distribution of the home types. The sentiment analysis revealed that sentiments and textual description do not have an impact on the house prices.

Next Steps

I would like to incorporate and explore the following additional features to enhance the predictive capacity of my project.

- **Larger Dataset:** I would like to gather more data particularly to address the limitations posed by the small number of single-family and townhouse properties in the current data. A larger dataset would provide better insights into the effect of property type on housing prices, enabling more robust modeling and reducing biases.
- **Time Series Analysis:** It will be helpful to incorporate time series data into analysis to capture market trends and seasonal variations in housing prices. This would allow for a more dynamic understanding of how prices evolve over time and could enable predictions about future prices based on historical patterns.
- **Expanded Feature Set:** Factors such as crime rates, accessibility to key infrastructures like schools, hospitals, and public transport, and proximity to amenities such as parks or shopping centers could significantly enhance the model's ability to explain price variability. Hence, it would be interesting to gather this information and include in the models.
- **Community Demographics and Economic Factors:** I feel like expanding the dataset to include socio-economic data about neighborhoods could provide deeper insights into how community attributes influence housing prices. Information about the racial and ethnic composition of the community, median income levels, and overall economic status of residents could offer valuable context for understanding price disparities across regions.

By integrating these features, I would like to not only improve the predictive performance of the model but also provide deeper insights into the socio-economic and temporal factors driving real estate trends.

References

[1] https://brightdata.com/cp/datasets/browse/gd_lfqkr8wm13ixtbd8f5?tab=sample&camp=plg

[2] https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524