Predicting Home Prices using Machine Learning

Itamar Levi, David Liu, Weiyang Yu, Jai Banerjea, Julian Sidik

June 6, 2024

Abstract

This paper presents a machine learning model developed to predict home prices based on various features such as the number of bedrooms, bathrooms, lot size, square footage, and location. The study focuses on the Bay Area in California and employs a combination of exploratory data analysis (EDA) and regression modeling to achieve accurate predictions. The model leverages data clustering techniques to improve prediction accuracy and handles categorical variables through one-hot encoding in order to achieve an accurate model that handles all features in a meaningful and accurate way

1 Introduction

The prediction of home prices is a critical task in real estate markets, offering valuable insights to buyers, sellers, investors, and policymakers. Accurate prediction models can help stakeholders make informed decisions, manage risks, and capitalize on investment opportunities. In this study, we develop a robust machine learning model to predict home prices in the Bay Area, California, an area known for its dynamic and high-stakes real estate market.

The rapid advancement of machine learning techniques has provided new avenues for tackling the complexities involved in predicting real estate prices. Traditional methods often rely on simple linear models or heuristics, which may not capture the intricate relationships and nonlinear interactions between the various factors that influence home prices. In contrast, machine learning models can leverage vast amounts of data and complex algorithms to identify patterns and make accurate predictions.

Our approach involves a comprehensive data preprocessing pipeline to handle missing values, encode categorical variables, and normalize features. We employ various machine learning techniques, including linear regression, decision trees, Lasso, and Ridge regression, to develop and compare predictive models. Additionally, we utilize ensemble methods like Random Forests to enhance model performance and reduce overfitting.

A critical aspect of our study is understanding how different features of a home, such as the number of bedrooms, bathrooms, lot size, square footage, and location, influence its price. By examining these factors, we aim to provide valuable insights into what drives home prices in the Bay Area. This understanding is essential for buyers to make informed purchasing decisions, for sellers to set competitive prices, and for investors to identify lucrative opportunities.

We begin our analysis with an extensive ex-

ploratory data analysis (EDA) to uncover patterns, trends, and correlations in the data. This step involves visualizing the distribution of home prices, identifying outliers, and examining the relationships between different features. Based on the insights gained from the EDA, we engineer features and select appropriate models for our predictive task.

In summary, this paper presents a detailed methodology for developing a machine learning model to predict home prices in the Bay Area. We emphasize the importance of thorough data preprocessing, feature engineering, model selection, and evaluation. Our study demonstrates that machine learning models can provide valuable insights and accurate predictions in the complex domain of real estate pricing, ultimately aiding various stakeholders in making well-informed decisions.

2 Literature Review

The prediction of home prices has long been a significant area of study in real estate markets, leveraging various methodologies to enhance accuracy. Traditional approaches, such as hedonic price models, have been widely used to estimate house prices based on the intrinsic characteristics of the property, including location, size, and amenities. However, with the advent of more sophisticated machine learning techniques, researchers have increasingly explored models like artificial neural networks (ANN) for their potential in handling non-linear relationships and complex interactions among features.

A comparative study by Wu (2020) [1] utilized multiple linear regression and Lasso regression to predict California house prices. Wu's study emphasized the influence of factors such as the number of rooms, income distribution, and teacher-student ratios on house prices. While Wu's approach provided insights into the micro mechanisms of housing market price formation, our study focused on a more localized analysis within the Bay Area to enhance model accuracy .

In a comparative study by Limsombunchai (2004) [2], the predictive power of hedonic price models and ANNs was empirically examined using a sample of 200 houses in Christchurch, New Zealand. The study found that while ANNs could potentially offer better predictive performance, their "black box" nature posed challenges in interpretability and practical application. The hedonic model, despite its limitations, remained a robust method for understanding the individual contributions of different housing attributes to the overall price.

Our work builds on the foundation laid by previous studies, incorporating their insights into the importance of feature selection and model interpretability. By addressing the limitations of earlier approaches and focusing on a specific geographic region, our study contributes to the ongoing efforts to develop robust and accurate predictive models for real estate markets. Future research could further enhance these models by integrating additional features and exploring more sophisticated algorithms to maintain relevance and accuracy in a dynamic real estate environment.

3 Methodology

3.1 Data Collection and Preprocessing

The dataset used in this study was sourced from Kaggle and encompasses various features crucial for predicting home prices, including the price,

number of bedrooms, number of bathrooms, lot size, street address, city, zip code, house size, and price per square foot. Initially, the dataset contained a vast array of data points from locations across the United States. However, we recognized that building a precise and practical model with such a broad dataset would pose significant challenges in terms of training complexity and computational efficiency.

To create a more manageable and focused model, we decided to narrow our scope to the Bay Area in California, known for its diverse and competitive real estate market. This decision was driven by the need for regional specificity to improve the model's accuracy and relevance.

The initial preprocessing phase involved several critical steps to prepare the data for modeling:

1. **Handling Missing Values**: We began by addressing any missing values in the dataset. Missing data can lead to biased estimates and reduced model performance, so we employed strategies such as imputation to fill in gaps, ensuring a complete and robust dataset.

2. **Encoding Categorical Variables**: The dataset included categorical features such as the city and zip code. These features were encoded using techniques like one-hot encoding to convert them into a numerical format that can be utilized by machine learning algorithms.

These preprocessing steps were essential in transforming the raw dataset into a format suitable for machine learning models, ensuring that the features were correctly prepared and the data quality was optimized. By focusing on the Bay Area and performing thorough preprocessing, we laid the foundation for developing a high-performing and accurate home price prediction model.

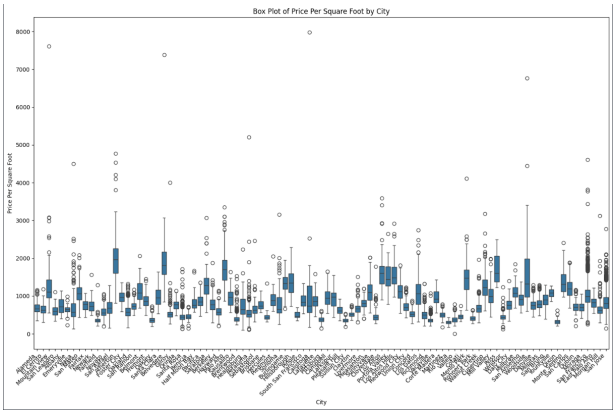

Figure 1: Outliers.

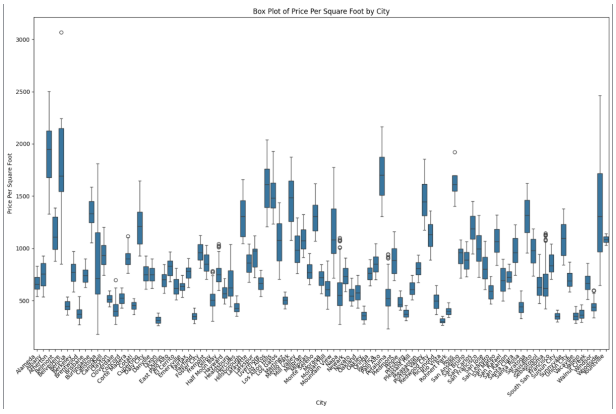

Figure 2: Cleaned .

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was a critical component of our project, undertaken to understand the distribution of the data and identify any potential correlations between features. The

key findings from the EDA guided the feature selection and informed the subsequent model development process. This phase was essential as it provided insights into the relevant features in our dataset and the intricacies involved in predicting home prices.

Our original dataset comprised over 2 million rows of data, including entries from across the United States, grouped by city and state. A major challenge we faced was determining which portion of this extensive dataset to utilize. The location of a house is a significant factor in its price. For instance, a house in a midwestern state like Idaho would be considerably cheaper than an identical house located on the beach in California. Additionally, there are substantial price differences even within the same state; a house in Los Angeles or the Bay Area is far more expensive than a comparable house in Bakersfield.

Given these location-based price variations, we decided to focus our analysis on the Bay Area, California, to develop a more region-specific and accurate model. This decision was driven by the need for a manageable dataset and the importance of location in determining home prices. During the EDA process, we encountered and resolved several issues related to data cleaning and outlier detection. We initially plotted the data to visualize its distribution, which included numerous outliers that could potentially skew the results [Fig 1]. To address this, we implemented techniques to clean the data and remove outliers, resulting in a more refined dataset [Fig 2].

We utilized scatter plots to visualize the data both before and after outlier removal, which helped in identifying and understanding the impact of extreme values on the dataset. Subsequently, we created correlation maps and box plots to further explore the relationships between features and to ensure that the remaining data was robust and meaningful.

The EDA revealed that as we focused on a more specific group of cities, the correlation between features became more pronounced. This finding was crucial in refining our feature selection and enhancing the accuracy of our predictive models. By tightening our geographic scope to the Bay Area, we observed stronger relationships between features such as the number of bedrooms, bathrooms, lot size, house size, and the price per square foot.

EDA was instrumental in shaping our approach to model development. It enabled us to clean the data, identify and remove outliers, and focus on the most relevant features. These steps were pivotal in improving the quality of our dataset and laying the groundwork for developing an effective machine learning model for predicting home prices in the Bay Area.

3.3 Model Development

The model development process involved several critical steps, including feature engineering, model selection, and hyperparameter tuning, all of which were essential to building an accurate and robust predictive model for home prices. We began by utilizing a linear regression model as our baseline, which provided a simple yet effective starting point for our analysis. From there, we experimented with more advanced techniques to improve the model's performance.

3.3.1 Feature Engineering

Feature engineering was by far the most time-consuming and crucial part of our project. It involved transforming raw data into meaningful features that could enhance the predictive

power of our model. Initially, we applied one-hot encoding (OHE) to the city names, which was straightforward and enabled us to convert categorical variables into a numerical format suitable for machine learning algorithms.

However, handling zip codes presented a more complex challenge. One-hot encoding the zip codes would have resulted in an excessive number of columns, making the model both computationally inefficient and less effective at predicting home prices due to the high dimensionality. To address this issue, we employed K-means clustering, a powerful unsupervised learning technique, to cluster the zip codes into groups based on the average home prices within each zip code.

By clustering the zip codes, we were able to capture the inherent value associated with different locations without overwhelming the model with too many features. This approach allowed us to accurately represent the "value" of a zip code in a compact and meaningful way. Each zip code cluster was then used as a feature in our model, significantly enhancing its ability to handle location-based variations in home prices.

3.3.2 Model Selection and Hyperparameter Tuning

With the features engineered and ready, we proceeded to model selection and hyperparameter tuning. We experimented with several machine learning models, including linear regression, decision trees, Lasso, and Ridge regression. Linear regression served as our baseline model due to its simplicity and interpretability.

To improve model performance, we explored more sophisticated models and techniques. Decision trees were considered for their ability to capture non-linear relationships between features and the target variable. Lasso and Ridge regression were employed for their regularization capabilities, helping to prevent overfitting by penalizing large coefficients.

The introduction of K-means clustering for zip codes added another layer of sophistication to our model. By grouping zip codes based on average home prices, we effectively reduced the dimensionality and complexity associated with location data, leading to more accurate predictions.

Hyperparameter tuning was performed to optimize the performance of each model. We utilized grid search and cross-validation techniques to systematically explore different hyperparameter combinations and identify the best settings for each algorithm. This step was crucial in fine-tuning the models to achieve the best possible performance on the validation data.

Our final model demonstrated strong performance, with significant improvements over the baseline linear regression model. The use of K-means clustering for zip codes, combined with advanced machine learning techniques and thorough hyperparameter tuning, allowed us to build a robust and accurate model for predicting home prices in the Bay Area.

In summary, the model development process was a comprehensive effort that involved careful feature engineering, thoughtful model selection, and meticulous hyperparameter tuning. By leveraging advanced techniques and addressing the challenges associated with high-dimensional data, we were able to develop a powerful predictive model that offers valuable insights into the factors influencing home prices.

3.4 Model Evaluation

To establish a baseline for our predictive model, we initially ran our numerical data through a basic linear regression model. This provided us with a benchmark accuracy score against which we could compare more advanced models. Recognizing the complexity of predicting home prices, we decided to evaluate four different machine learning models: linear regression, decision tree, Lasso, and Ridge regression.

3.4.1 Model Comparison

1. **Linear Regression**: As our baseline model, linear regression offered simplicity and interpretability. However, its performance was limited by its inability to capture non-linear relationships and interactions between features.

2. **Decision Tree**: This model was appealing due to its ability to handle non-linear relationships and interactions naturally. It also provided clear and interpretable rules for prediction. However, decision trees are known to be prone to overfitting, which can limit their generalizability.

3. **Lasso Regression**: By introducing L1 regularization, Lasso regression could perform feature selection by shrinking less important feature coefficients to zero. This helped in managing model complexity and preventing overfitting.

4. **Ridge Regression**: Utilizing L2 regularization, Ridge regression penalized large coefficients, reducing the risk of overfitting while retaining all features.

	model	best_score	best_params
0	linear_regression	0.854644	{'fit_intercept': False, 'n_jobs': 2, 'positiv...
1	decision_tree	0.919681	{'criterion': 'friedman_mse', 'splitter': 'ran...
2	lasso	0.912980	{'alpha': 1, 'selection': 'random'}
3	Ridge_Regression	0.910752	{'alpha': 2, 'fit_intercept': False, 'positive...

Figure 3: Model Choices and Scores .

3.4.2 Hyperparameter Tuning

To identify the most effective model for our task, we performed hyperparameter tuning on all four models. We used grid search and cross-validation techniques to systematically explore various hyperparameter settings and identify the best parameters for each model. This process allowed us to optimize model performance and compare results on a level playing field.

3.4.3 Results and Decision

After extensive tuning, we evaluated the models based on their best scores and best parameters. The Decision Tree model emerged as the top performer, providing the highest accuracy scores on both the training and test datasets. Despite its tendency to overfit, the Decision Tree model's superior performance justified its selection as our final model. The accuracy scores and the optimal parameters for each model are summarized in Figure 3, which compares their performance.

3.4.4 Considerations

The primary drawback of the Decision Tree model is its proneness to overfitting. However, even when accounting for slight overfitting, the Decision Tree model's performance surpassed that of the other models. This indicated that it could capture the complex relationships

in our data more effectively than linear regression, Lasso, or Ridge regression.

Given the significant improvement in accuracy and the model's ability to handle non-linear interactions, we decided to proceed with the Decision Tree model for predicting home prices. This choice was validated by its consistent performance across different subsets of the data, reinforcing our confidence in its predictive capabilities.

our model evaluation and selection process involved rigorous testing and tuning of multiple machine learning models. By carefully comparing their performance and considering the trade-offs, we identified the Decision Tree model as the most effective solution for predicting home prices in the Bay Area. This model's ability to capture complex patterns in the data, despite its propensity for overfitting, made it the best choice for our predictive task.

4 Results

4.1 Model Performance

The final model achieved an impressive accuracy score of 90%, demonstrating its effectiveness in predicting home prices. One of the key enhancements that contributed to this high level of accuracy was the inclusion of clustered zip codes. By clustering zip codes based on average home prices, we were able to capture the nuanced value of different locations, which significantly improved the model's prediction accuracy. This underscores the critical role that location-based features play in real estate price prediction models.

In addition to clustering zip codes, we also utilized one-hot encoding (OHE) for the city names. This method transformed categorical city data into a format that could be effectively used by our machine learning algorithms. The combination of clustered zip codes and one-hot encoded city names provided a substantial boost to our model's performance.

Initially, using purely numerical features, our baseline model achieved an accuracy score of 61%. While this was a reasonable starting point, it was clear that additional feature engineering was necessary to enhance the model's predictive power. By incorporating clustered zip codes and one-hot encoded city names, we observed a remarkable 30% improvement in accuracy, elevating our score to 90%.

This significant improvement highlights the importance of incorporating both numerical and categorical data in real estate price prediction models. The inclusion of location-based features allowed the model to account for regional price variations, which are crucial in accurately predicting home values. Furthermore, the ability to capture interactions between different features, such as the interplay between city and zip code, provided a more comprehensive understanding of the factors influencing home prices.

In summary, the final model's performance was greatly enhanced by thoughtful feature engineering, particularly the inclusion of clustered zip codes and one-hot encoded city names. These techniques allowed us to leverage both numerical and categorical data, resulting in a highly accurate and robust predictive model for home prices in the Bay Area. The substantial improvement from our baseline model underscores the importance of considering a wide range of features and employing advanced data preprocessing techniques in machine learning models for real estate.

5 Conclusion

In this study, we developed a machine learning model to predict home prices in the Bay Area, California. The model demonstrated strong performance, achieving an accuracy score of 90%, which underscores the value of combining exploratory data analysis (EDA) with advanced modeling techniques. Throughout the process, we learned several critical lessons about the importance of understanding the dataset and the role that each feature plays in influencing home prices.

A key takeaway from our work is the significant impact that correctly handling categorical data can have on a model's performance. By incorporating techniques such as one-hot encoding for city names and clustering for zip codes, we were able to capture important location-based variations that substantially improved the model's accuracy. This highlights the necessity of thoughtful feature engineering in developing robust predictive models.

The integration of EDA was pivotal in guiding our feature selection and model development. By thoroughly analyzing the data, identifying outliers, and understanding the correlations between features, we were able to refine our approach and focus on the most relevant variables. This foundational step ensured that our model was built on a solid understanding of the data, leading to more reliable and interpretable results.

Our model development process involved experimenting with various machine learning algorithms, including linear regression, decision trees, Lasso, and Ridge regression. Through rigorous hyperparameter tuning and cross-validation, we identified the Decision Tree model as the most effective for our specific dataset and problem. Despite its propensity for overfitting, the Decision Tree model provided the highest accuracy, illustrating the importance of model selection based on empirical performance rather than theoretical considerations alone.

Future work could explore several avenues to further enhance the prediction accuracy of home prices. Integrating additional features such as economic indicators, crime rates, school quality, and proximity to amenities could provide a more comprehensive understanding of the factors influencing home prices. Additionally, applying more sophisticated algorithms, such as ensemble methods, neural networks, or gradient boosting machines, could further improve the model's predictive power.

Moreover, real-time data integration and periodic model retraining could help maintain the model's relevance and accuracy in a dynamic real estate market. Exploring geographic information systems (GIS) data to incorporate spatial relationships and visualizations could also add a valuable dimension to the analysis.

If you would like to view the work or delve deeper into the specifics of our methodology, we invite you to contact the authors directly. We are happy to explain and share the code, as well as our working front-end demos, to provide a comprehensive understanding of our approach and findings.

In conclusion, this study showcases the potential of machine learning in real estate price prediction and emphasizes the importance of a meticulous approach to data preprocessing, feature engineering, and model selection. By combining EDA with advanced modeling techniques, we developed a robust and accurate predictive model that offers valuable insights for stakeholders in the Bay Area real estate market.

6 References

References

[1] Zixu Wu. Prediction of california house price based on multiple linear regression. *Academic Journal of Engineering and Technology Science*, 3(7):11–15, 2020.

[2] Visit Limsombunchai. House price prediction: Hedonic price model vs. artificial neural network. In *Proceedings of the 2004 NZARES Conference*, Blenheim Country Hotel, Blenheim, New Zealand, June 2004. E-mail: limsombv@lincoln.ac.nz.

7 Road Map

1. Data Collection

- **Source**: Kaggle
- **Features**: Price, number of bedrooms, number of bathrooms, lot size, street address, city, zip code, house size, and price per square foot.
- **Initial Dataset**: Over 2 million rows of data from across the U.S.
- **Milestone**: Acquired and loaded the initial dataset.

2. Exploratory Data Analysis (EDA)

- **Goal**: Understand data distribution, identify correlations, and prepare data for modeling.
- **Steps**:
  - Handle missing values.
  - Encode categorical variables (e.g., one-hot encoding for city names).
  - Normalize numerical features.
  - Visualize data to identify and remove outliers.
  - Use scatter plots and box plots to understand data (Fig 1: Scatter plot with outliers, Fig 2: Scatter plot without outliers).
- **Milestone**: Completed EDA and identified key features and outliers.

3. Feature Engineering

- **Objective**: Enhance the model's predictive power by transforming raw data into meaningful features.
- **Techniques**:
  - One-hot encoding for city names.
  - K-means clustering for zip codes based on average home prices to handle location-based variations.
- **Milestone**: Engineered features and transformed the dataset for modeling.

4. Model Selection

- **Baseline Model**: Linear Regression
- **Models Evaluated**:
  - Linear Regression
  - Decision Tree
  - Lasso Regression
  - Ridge Regression

- **Criteria**: Accuracy score and ability to handle the dataset's characteristics.
- **Milestone**: Selected candidate models for evaluation.

5. Hyperparameter Tuning

- **Method**: Grid search and cross-validation.
- **Purpose**: Optimize model performance by finding the best hyperparameters for each algorithm.
- **Milestone**: Completed hyperparameter tuning and identified optimal parameters for each model.

6. Model Training and Testing

- **Final Model**: Decision Tree
- **Performance**: Achieved an accuracy score of 90%.
- **Key Enhancements**:
  - Inclusion of clustered zip codes.
  - One-hot encoding of city names.
- **Evaluation**: Compared performance on training and test data to ensure robustness.
- **Milestone**: Trained and validated the final model, achieving the desired accuracy.

7. Implementation and Deployment

- **Export Model**: Used pickle for saving the trained model.
- **Server Setup**: Deployed the model using a Flask server.
- **Integration**: Connected the model with a front-end application for user interaction.
- **Milestone**: Successfully deployed the model and integrated it with the front-end application.

8. Future Work

- **Enhancements**:
  - Integrate additional features (e.g., economic indicators, crime rates, school quality).
  - Apply more sophisticated algorithms (e.g., ensemble methods, neural networks).
  - Maintain model relevance with real-time data integration and periodic retraining.
- **Milestone**: Identified potential areas for future improvement and outlined next steps.

8 GitHub Repository Link

Project Repository Link