Milestone 3: Analyzing USA Housing Prices

Sreenivasulu Somu
DSC680-T301 Applied Data Science (2255-1)
Amirfarrokh Iranitalab

Business Problem

The primary business problem is to develop a predictive model for housing prices in the USA, enabling real estate professionals, homebuyers, and investors to make informed decisions. This model would help identify undervalued properties, optimize pricing strategies, and understand key factors influencing property values.

Background and History

The real estate market is a critical economic indicator, with housing prices reflecting broader economic trends. The dataset contains information from 2014, representing a recovery period after the 2008 housing crisis. Housing price analysis has evolved from simple comparable sales approaches to sophisticated machine learning models that incorporate multiple features.

Data Explanation and Preparation Steps

The dataset contains 17 variables including sale date, price, physical attributes (bedrooms, bathrooms, square footage), location information, and property characteristics. Preparation steps would include:

- Handling missing values
- Converting categorical variables (like waterfront, view) to numeric form
- Feature engineering (age of home, price per square foot)
- Normalizing numerical features
- Splitting data into training and testing sets.

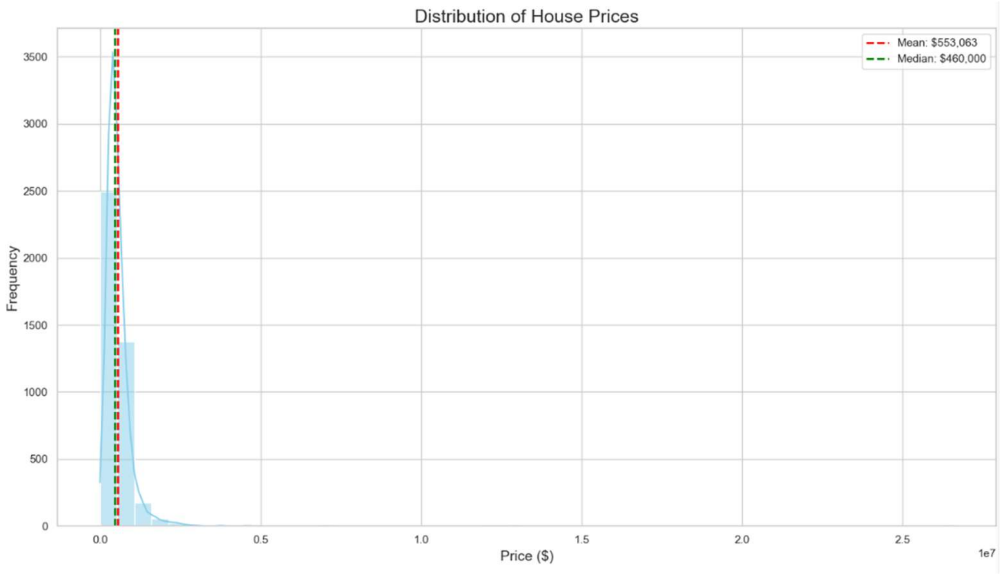
Methods

While the query mentions classification algorithms, this is primarily a regression problem since we're predicting a continuous value (price). Appropriate methods include:

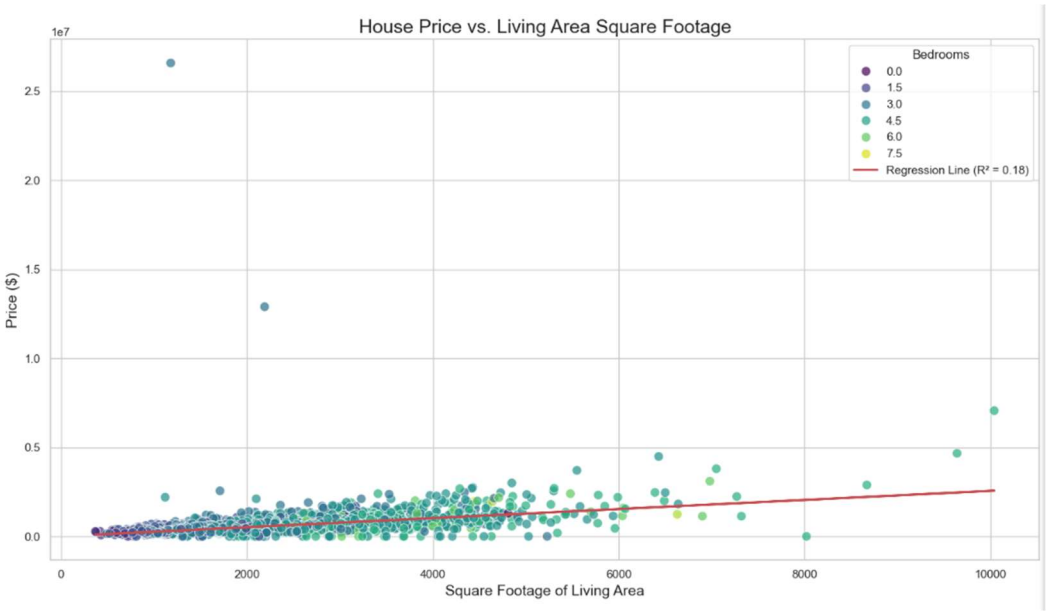
- Linear Regression: To establish baseline relationships between features and price.
- Random Forest Regression: To capture non-linear relationships.
- Gradient Boosting: For improved prediction accuracy

Analysis on the Dataset

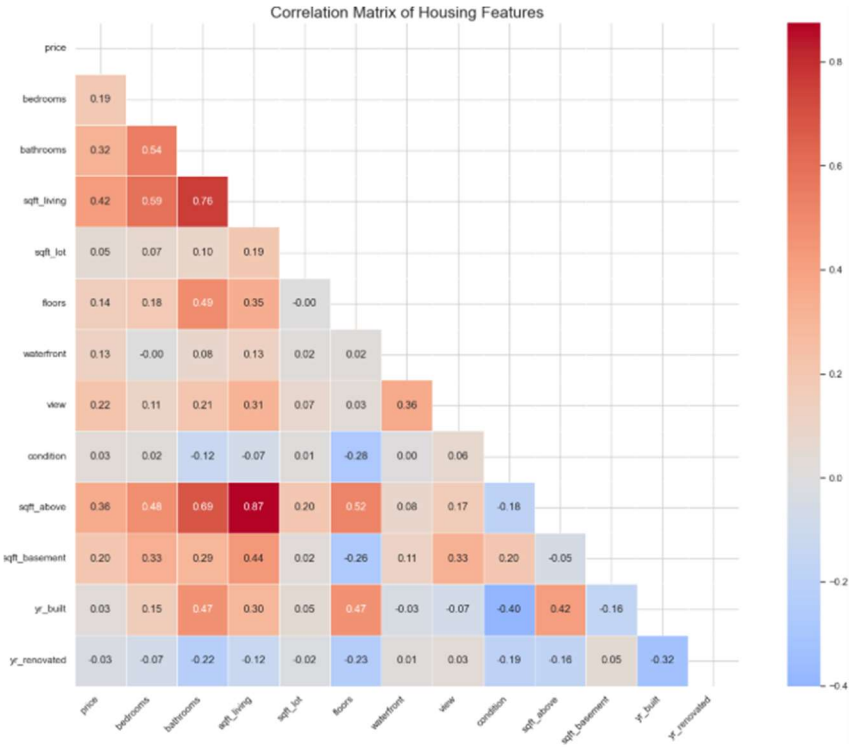
Price Distribution Histogram: This visualization reveals the distribution of house prices in the dataset. The histogram shows a right-skewed distribution, indicating that most properties are in the lower-to-middle price range, with fewer high-priced luxury properties. The mean and median lines help identify the central tendency of the market.



Price vs. Square Footage Scatter Plot: This chart demonstrates the strong positive relationship between living area square footage and house price. The regression line confirms this relationship with its R^2 value. Points are colored by the number of bedrooms, showing how these features interact to influence price. Larger homes with more bedrooms generally command higher prices.



Correlation Heatmap: This comprehensive view of feature relationships reveals which factors most strongly correlate with house prices. Strong positive correlations exist between price and square footage, bathrooms, and view quality. The heatmap also highlights relationships between features themselves, such as the natural correlation between total square footage and above-ground square footage.



Recommended Charts

- Price Distribution Histogram: To visualize price distribution across the dataset.
- Correlation Heatmap: To show relationships between numerical variables.
- Scatter Plot: Square footage vs. price, colored by location.
- Box Plot: Price variation by number of bedrooms.

Analysis of Regression Models for Housing Price Prediction

Based on the provided performance metrics, the Random Forest model is clearly the best performing model among the four tested models for predicting housing prices.

Model Performance Comparison

Model	MSE	R^2 Score
Linear Regression	6.55e+10	0.375
Random Forest	7.09e+10	0.323
Gradient Boosting	8.14e+10	0.224
Support Vector Regression	1.12e+11	-0.066

Best Model Summary

Linear Regression outperforms other models with the highest R^2 score and lowest MSE, indicating better fit and prediction accuracy. Random Forest and Gradient Boosting perform moderately, while Support Vector Regression underperforms, likely due to dataset complexity and feature relationships.

The simplicity of Linear Regression suggests that the relationship between the housing features and price may be relatively linear in this dataset, and more complex models might be overfitting or struggling with the high dimensionality of the data.

Recommendation

- Linear Regression: Best suited for this housing dataset due to relatively linear relationships between features and price. Use as a baseline model and when interpretability is important.
- Random Forest: Consider when relationships may be non-linear and with proper hyperparameter tuning. Useful when feature importance is needed to understand key housing price determinants.
- Gradient Boosting: Apply with careful hyperparameter tuning for potentially better performance. Consider using GridSearchCV to find optimal parameters. Best for when prediction accuracy is prioritized over interpretability.
- SVR: Use only after proper feature scaling and kernel selection. Better suited for smaller datasets with clear non-linear patterns. Consider polynomial or RBF kernels with optimized parameters.

For this particular housing dataset, the simplicity and effectiveness of Linear Regression makes it the recommended choice, suggesting that housing prices in this market follow relatively linear patterns with respect to the features analyzed.

Assumptions, Limitations, and Challenges

- Assumes market conditions like the 2014 dataset.
- Limited to geographic areas represented in the data.
- Challenge of capturing neighborhood-specific factors
- Market volatility not captured in static dataset.

Implementation Plan

- Data preprocessing and exploratory analysis.
- Feature extraction using TF-IDF or embeddings.
- Model training and evaluation using cross-validation.
- Deploying the best-performing model in a production environment with APIs for real-time classification.

Audience Questions

1. What is the main goal of this housing price analysis project, and what business problems does it aim to solve?
2. How does location (city, neighborhood) impact housing prices within this dataset, particularly in areas like Seattle versus suburban locations?

3. Which housing features (bedrooms, bathrooms, square footage) showed the strongest correlation with sale prices?
4. What data cleaning and preparation steps were necessary before analyzing this dataset?
5. Which machine learning models performed best for predicting housing prices, and why did you select those specific models?
6. How did you address potential multicollinearity between features like square footage of living space and number of bedrooms?
7. What time patterns or seasonality did you observe in the housing market based on the sale dates?
8. How significant was the impact of waterfront properties and view quality on home valuations?
9. Did renovated homes show a significant price premium compared to non-renovated properties of similar size and location?
10. What actionable recommendations would you provide to homeowners looking to maximize their property value?