Predicting House Prices in King County, USA: A Data-Driven Analysis

Gulnar Armour
Springboard
December 26, 2023

Introduction

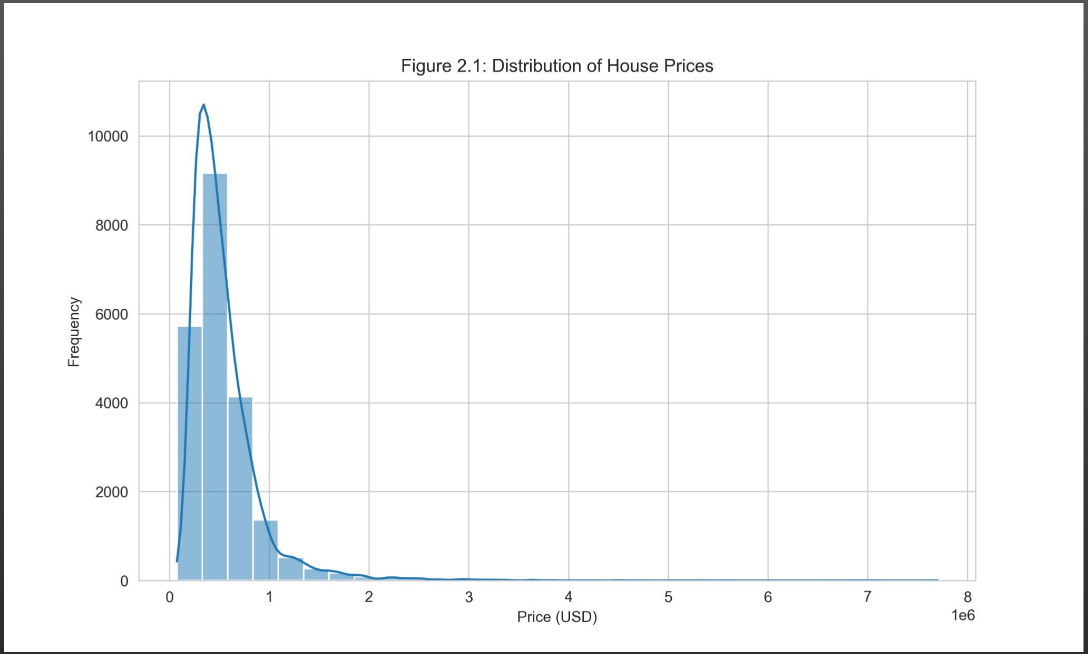
- Project Overview:
  - Utilization of machine learning techniques to predict house prices
  - Data source: "House Sales in King County, USA" dataset from Kaggle
- Project Focus:
  - Building and evaluating machine learning models for accurate house price forecasting
  - Providing comprehensive insights for real estate stakeholders and prospective homebuyers
- Primary Objective:
  - Development of robust machine learning models for precise house price prediction

Dataset Description

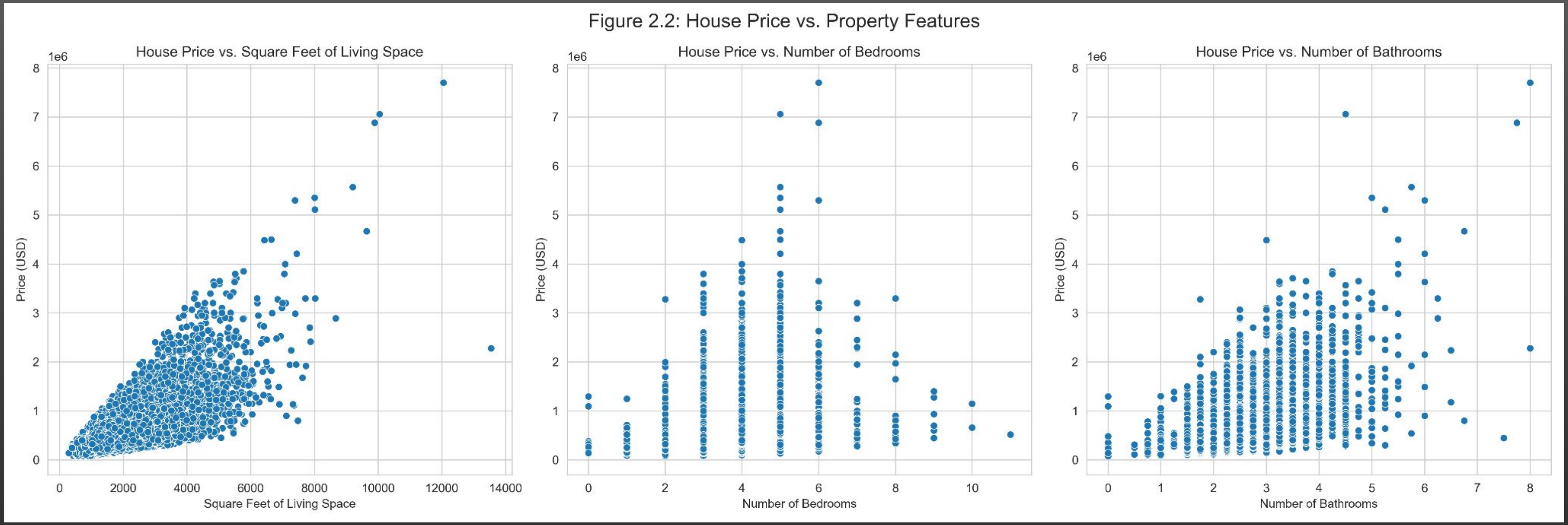
- Dataset overview:
  - 21,613 records
  - 20 features
- Attributes:
  - Property characteristics
  - Location data
  - Comparative transaction information
- Mix of numerical and categorical features for analyzing and predicting house prices in King County, USA

Univariate Analysis of House Prices

- Histogram showing the distribution of house prices in the dataset
- Nature of Distribution: **right-skewed**
- Implication: majority of houses are priced lower, with fewer houses in the higher price range
- Typicality: reflects a common distribution pattern for house prices
- Importance: helps in understanding the range and variance of house prices in the dataset



Bivariate Analysis: House Price vs. Key Features



- **House Price vs. Square Feet of Living Space (graph: 1):** positive correlation, larger homes typically have higher prices
- Trend: increase in living space often leads to an increase in house price
- **House Price vs. Number of Bedrooms (graph: 2):** positive correlation but with more variability than living space
- Observation: more bedrooms usually mean higher prices, but the relationship is less linear
- **House Price vs. Number of Bathrooms (graph: 3):** positive correlation noted between the number of bathrooms and house prices
- Variability: Significant variability in this relationship

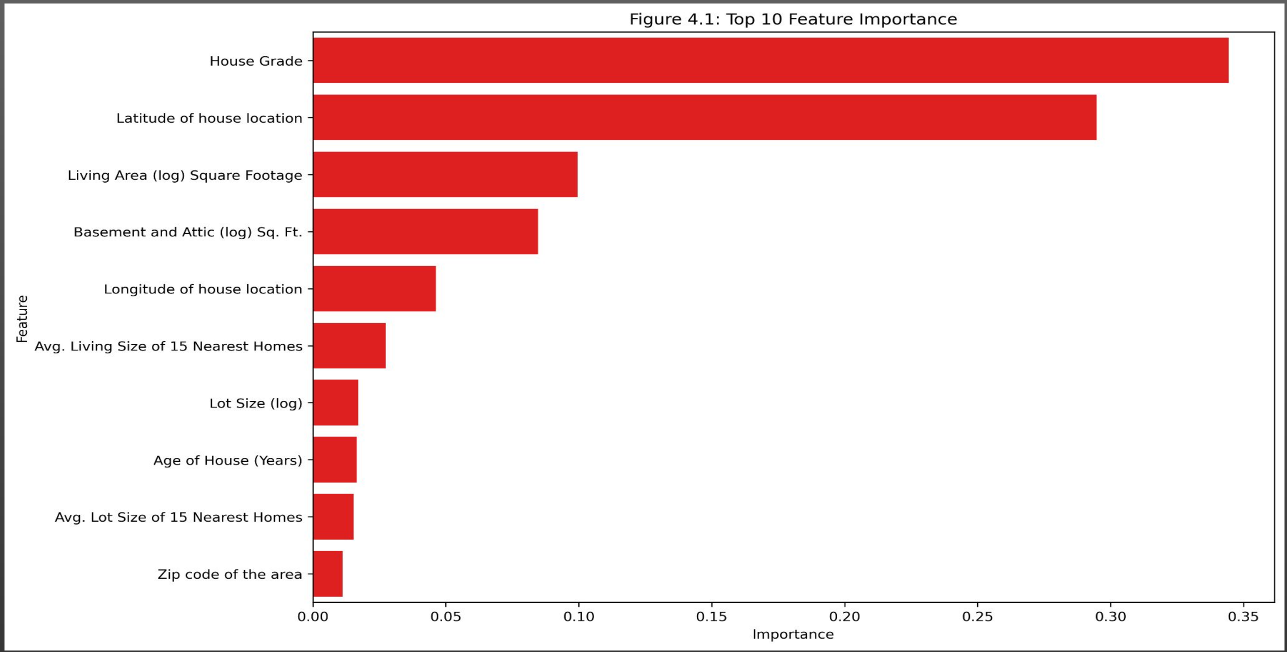
Statistical Analysis

Table 2.1: The Pearson correlation

Feature	Correlation Coefficient	P-Value	Result
Square Feet of Living Space	0.702	<0.001	Statistically Significant
Number of Bedrooms	0.315	<0.001	Statistically Significant
Number of Bathrooms	0.525	<0.001	Statistically Significant

- **Correlation between Living Space and House Price:**
  - Strong correlation (0.702) found
  - Larger living spaces are typically associated with higher property values
- **Correlation between Number of Bedrooms and House Price:**
  - Positive but weaker correlation (0.315) observed
  - Bedrooms influence property prices, yet other factors also contribute significantly
- **Correlation between Number of Bathrooms and House Price:**
  - Moderately strong correlation (0.525) noted
  - More bathrooms usually correlate with higher property values
- **Implications of Findings:**
  - Emphasizes the importance of living space, bedrooms, and bathrooms in influencing house prices
  - Provides valuable insights for real estate professionals and market stakeholders

Feature importances



- Top 10 features impacting house prices in King County, USA, as per Random Forest model analysis.
- Most Influential Feature - House Grade:
  - Significance: reflects overall quality and construction standards
  - Insight: indicates high buyer emphasis on build quality
- Second Most Critical Factor - Latitude of House Location:
  - Importance: north-south positioning of the property
  - Impact: plays a pivotal role in market value determination

Predictive Modeling

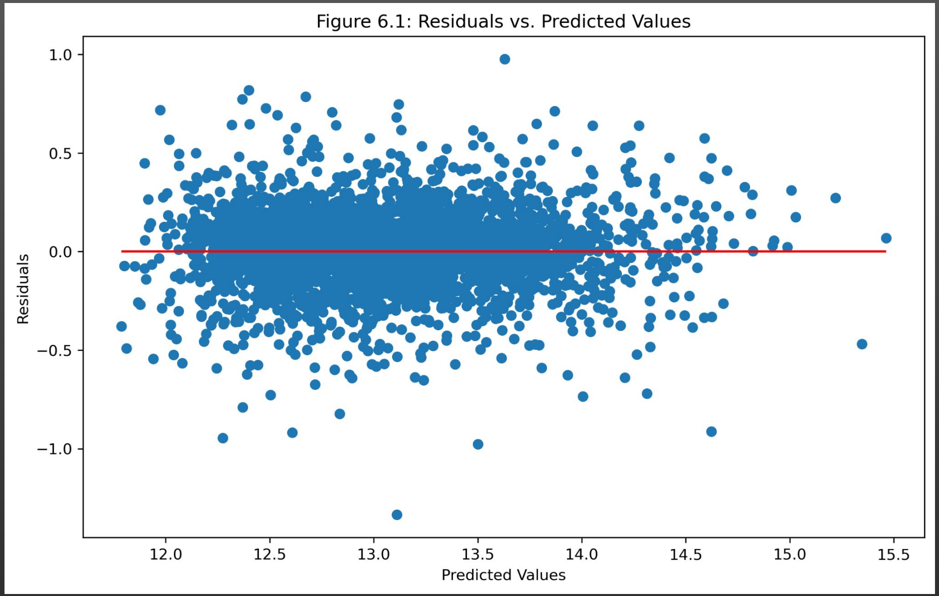
Table 3.1: The Root Mean Square Error (RMSE) Values

Model	RMSE with Non-Logged Features	RMSE with Logged Features
Linear Regression	212,242.35	198,835.51
Random Forest	147,732.82	**139,035.95**
Decision Tree	206,839.73	183,689.14
Gradient Boosting	149,466.78	145,788.72

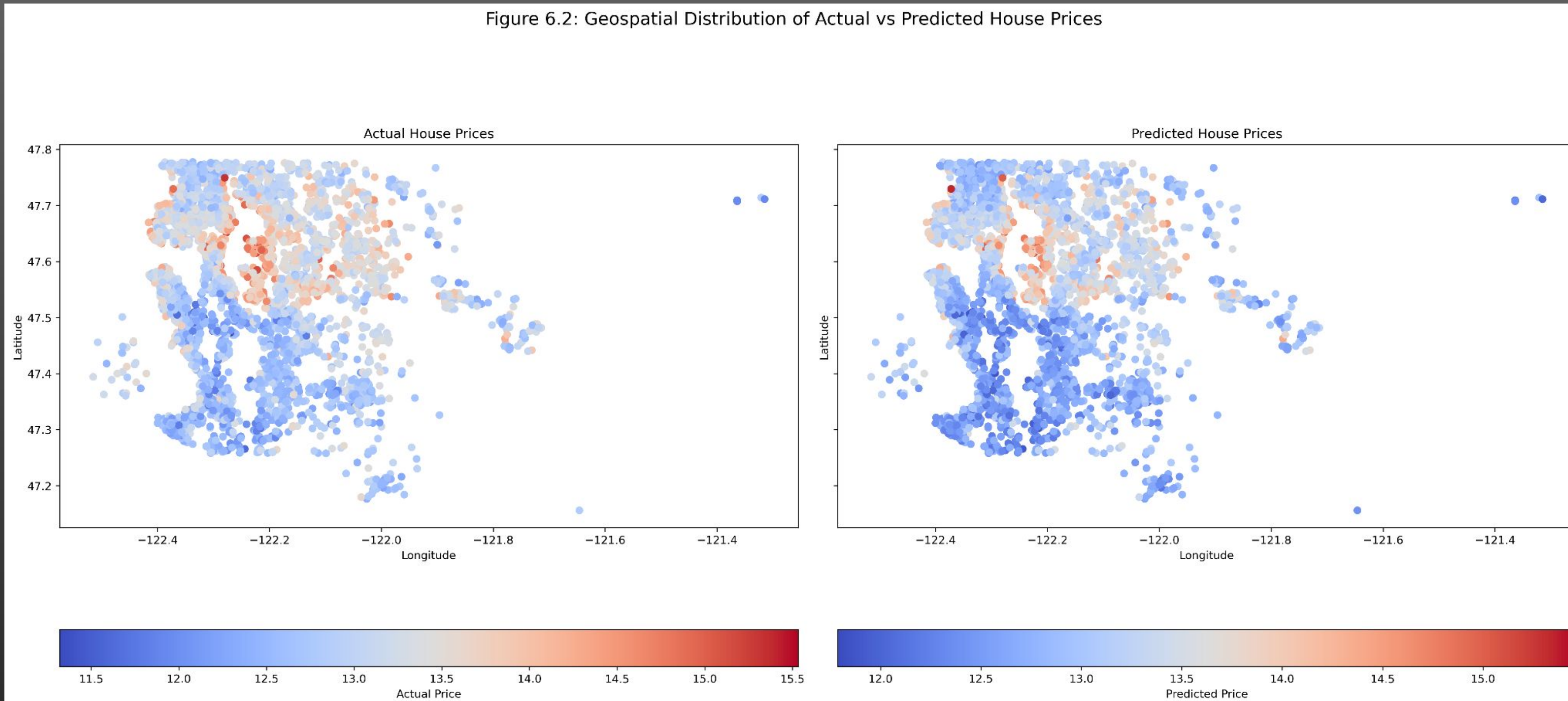
- **Model Selection Objective:** identify a model that provides accurate predictions and effectively handles data characteristics
- **Evaluation Criteria:**
  - Method: assessment of various models using both non-logarithmic and logarithmic features
  - Key Metric: Root Mean Square Error (RMSE) as the primary evaluation standard
- **Outcome of Comparative Analysis:**
  - Random Forest model emerged as superior for accurately predicting house prices
  - Effective handling of both non-logarithmic and logarithmic features; lowest RMSE scores
- **Insight on Data Normalization:**
  - RMSE improvement with logarithmic features across all models
  - Normalizing data distributions enhances model performance, confirming the initial hypothesis

Residual Analysis

- Predictions are neither systematically overestimated nor underestimated
- Suggests overall accurate performance of the model
- Slight increase in spread of residuals for higher predicted prices
- Predictions become less precise for more expensive houses
- Outliers and Leverage Points:
  - Presence of points straying far from the zero line, especially at higher predicted values
- Potential Causes:
  - Unique property features, emerging market trends, or data entry errors



Geospatial Validation



- Actual Prices: Intense reds for high-value properties amidst cooler blues.
- Predicted Prices: Remarkable visual congruence with the actual price distribution.
- Implication: Model effectively captures geographic nuances in property valuation.

Conclusion

- Model Selection and Evaluation:
  - Discovery: Random Forest model as the most accurate among tested algorithms
  - Validation: Cross-validation and residual analysis affirming model reliability
- Key Insights from Feature Analysis:
  - House grade and latitude identified as pivotal in determining house prices
  - Implications: highlights the importance of regional components and location in property valuation
- Geospatial Validation Insights:
  - Visualized model's prediction accuracy across different locations
  - Confirmed model effectiveness and provided real-world applicability of the data
- Summary of Findings:
  - Advanced machine learning techniques are invaluable in real estate market analysis
  - Quality, location, and size as key factors in property valuation
- Future Directions:
  - Exploring additional features and more sophisticated models
  - Expanding dataset to include temporal aspects for dynamic analysis