Predicting Housing Prices in King County using Machine Learning

Northeastern University

Course: ALY 6140: Python and Analytics Systems Technology

Professor Richard He

October 24, 2025

Submitted by:

Prasanthi Chitturi

Tharun Pallela

Meghana Yalam

Introduction

The housing market is one of the most dynamic and data-rich areas for undertaking predictive modeling and offers clear advantages for understanding the main drivers of property values and anticipating price behavior. This capstone develops a solid predictive framework for estimating residential housing prices in King County, Washington, which includes the city of Seattle and its surrounding suburb. Leveraging state-of-the-art machine learning algorithms, this study will shed light on the complexity and nuances of the relationship between structural, locational, and qualitative features that impact housing prices, as well as evaluate which modeling approach yielded the highest quality predictions. The dataset used is the "King County House Sales" data provided by Kaggle, which includes over 21,560 promissory sale transactions containing key features, such as living area, construction grade (numeric), number of bedrooms and bathrooms, waterfront, and latitude and longitude coordinates. These features capture both physical and non-physical drivers of property value. A key component of the analysis is exploratory data analysis (EDA) which uncovers trends, identifies outliers, and transforms the data for readiness for modeling. Features derived, such as house age and log-transformed price, improve the dataset capacity to predict housing prices, and also correct the skewness of the price distribution. Three primary supervised regression approaches - **Linear Regression, Random Forest Regression,** and **Extreme Gradient Boosting (XGBoost)** - were executed, each representing a different level of model complexity. Linear Regression is used as the interpretable benchmark, whereas ensemble-based algorithms (**Random Forest and XGBoost)** aim to capture non-linear relationships and high-order interactions that traditional methods often overlook. Model performance followed by R^2 and **Root Mean Square Error (RMSE)** is assessed with cross-validation to ensure replicability. The main research questions underpinning this study center on identifying the most significant predictors of housing prices, examining how well advanced models perform compared to classic models, and exploring the trade-off between interpretability and predictive power. Ultimately, this project hopes to demonstrate how machine learning can not only provide an accurate prediction of prices but also can provide interpretable information to support investment in real estate, policy, and market changes.

Proposed Research Questions

This project investigates several key questions related to housing price prediction and market dynamics in King County, Washington. These questions guided the design of the exploratory analysis and the selection of predictive models:

1. Which features most strongly influence house prices in King County?

   This explores how structural (e.g., square footage, bedrooms, bathrooms), qualitative (e.g., grade, condition), and locational (e.g., latitude, waterfront view) variables contribute to property valuation.
2. Can advanced machine learning methods improve the accuracy of housing price predictions compared to traditional linear models?

   This evaluates whether ensemble models such as Random Forest and XGBoost outperform Linear Regression in predictive performance.
3. How do geographic and quality-related variables interact with structural features in determining price?

   This examines whether non-linear interactions and complex dependencies are better captured through ensemble approaches.
4. Which model provides the best balance between interpretability and predictive accuracy?

   This question assesses model explainability versus performance trade-offs, identifying the optimal method for real-estate valuation tasks.

Exploratory Data Analysis (EDA)

1.Dataset Overview and Extraction

The dataset used in this study, King County House Sales, contains 21,560 property transactions and 25 variables, covering residential sales in Seattle and nearby suburbs between May 2014 and May 2015. Each observation represents a single property sale and includes key structural, locational, and qualitative attributes such as:

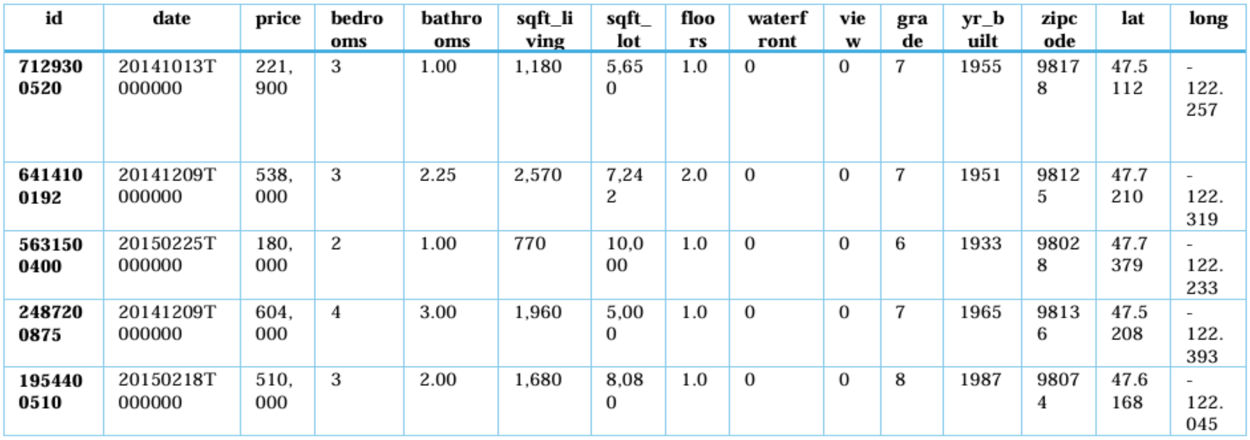
- Price – the sale price of the home (target variable)
- Bedrooms, Bathrooms, Floors – interior structural details
- Sqft_living, Sqft_lot, Sqft_above, Sqft_basement – measures of property size.
- Waterfront, View, Condition, Grade – qualitative indicators of property quality and location desirability.
- Year built and Year renovated – temporal attributes representing property age.
- Latitude, Longitude, and Zip code – geographical coordinates identifying location.

The dataset was imported using pandas_read_csv() and inspected through .head(), .info(), and .describe() to confirm structure and data types. There were no missing values across the dataset, ensuring reliability for modeling.

2. Data Cleaning and Feature Engineering

To improve data consistency and enhance model accuracy, several preprocessing steps were performed:

- **Date Parsing:** The date column was converted to datetime format, and new temporal features sale_year and sale_month were extracted for seasonal trend analysis.
- **Feature Creation:** A new variable house_age was computed as the difference between the year of sale and the year built, representing the age of the property. Additionally, price_log was created using a natural log transformation of price to correct the strong right-skewed distribution observed in the raw price data.
- **Outlier Removal:** Extreme outliers were filtered to ensure more stable model performance. Specifically:



id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	grade	yr_built	zipcode	lat	long
7129300520	20141013T000000	221,900	3	1.00	1,180	5,650	1.0	0	0	7	1955	98178	47.5112	-122.257
6414100192	20141209T000000	538,000	3	2.25	2,570	7,242	2.0	0	0	7	1951	98125	47.7210	-122.319
5631500400	20150225T000000	180,000	2	1.00	770	10,000	1.0	0	0	6	1933	98028	47.7379	-122.233
2487200875	20141209T000000	604,000	4	3.00	1,960	5,000	1.0	0	0	7	1965	98136	47.5208	-122.393
1954400510	20150218T000000	510,000	3	2.00	1,680	8,080	1.0	0	0	8	1987	98074	47.6168	-122.045

- Houses priced above $3,000,000 were removed (rare luxury outliers).
- Properties with > 10 bedrooms or > 6 bathrooms were excluded.
- Properties with more than 3.5 floors were also trimmed.

These filtering and transformation steps improved normality, reduced model bias, and maintained a balanced sample of typical residential properties.

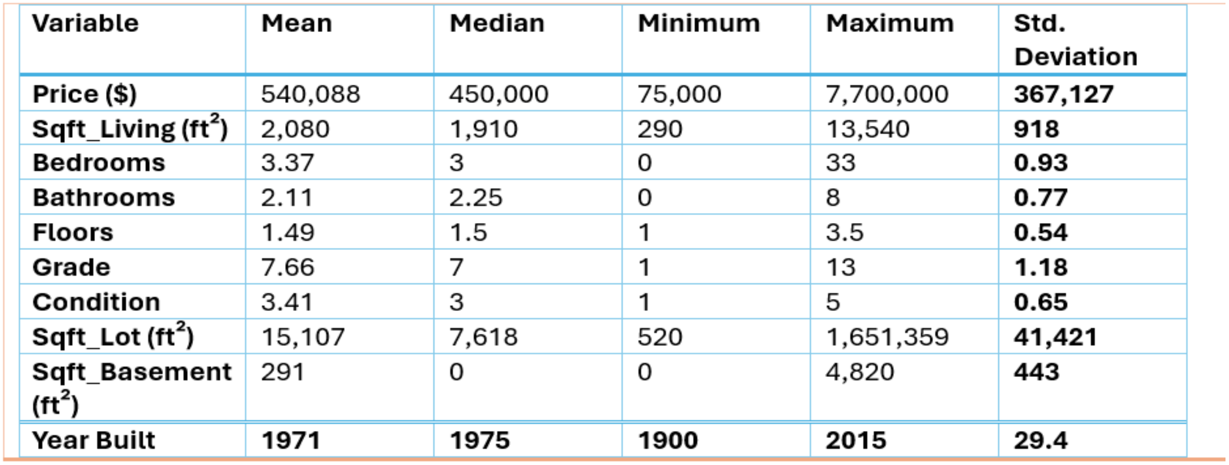
1. Descriptive Statistics

The descriptive summary shows that:

- The average home price is approximately $ 532,743, with values ranging from $75,000 to $3,000,000.
- The average living space (sqft_living) is about 2,070 square feet, with homes having three bedrooms and two bathrooms on average.
- The average property age is 44 years, and the median sale year is 2014.

These results reflect a diverse housing market with a wide range of property sizes and prices, typical of metropolitan regions like Seattle.

Table 1: Descriptive Statistics for Key Housing Variables (King County Dataset)



Variable	Mean	Median	Minimum	Maximum	Std. Deviation
Price ($)	540,088	450,000	75,000	7,700,000	**367,127**
Sqft_Living (ft^2)	2,080	1,910	290	13,540	**918**
Bedrooms	3.37	3	0	33	**0.93**
Bathrooms	2.11	2.25	0	8	**0.77**
Floors	1.49	1.5	1	3.5	**0.54**
Grade	7.66	7	1	13	**1.18**
Condition	3.41	3	1	5	**0.65**
Sqft_Lot (ft^2)	15,107	7,618	520	1,651,359	**41,421**
Sqft_Basement (ft^2)	291	0	0	4,820	**443**
Year Built	**1971**	**1975**	**1900**	**2015**	**29.4**

2. Data Visualization and Interpretation

Visual analysis was conducted to identify relationships and data patterns

a. House Price Distribution

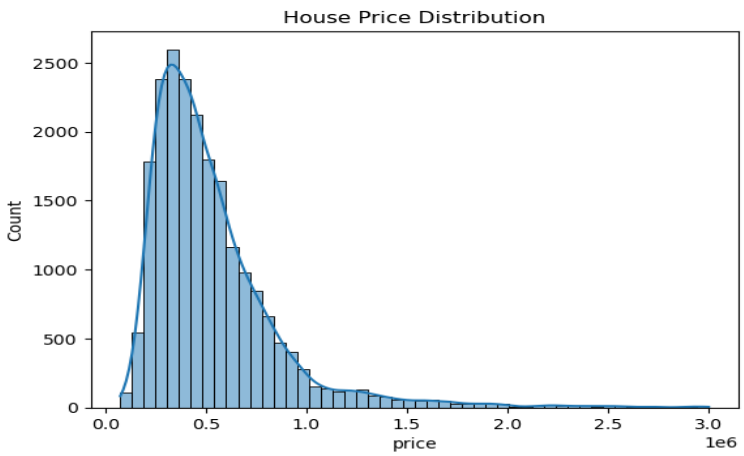


Figure 1: House Price Distribution

The distribution of prices is heavily right-skewed, indicating that most homes are moderately priced while a few luxury properties pull the average upward. The log transformation (price_log) was applied to stabilize this skewness.

b. Correlation Analysis

The correlation heatmap shows strong relationship between price and several features: Sqft_living (r = 0.69); Grade (r = 0.68); Sqft_above (r = 0.60).

These suggest that larger homes and higher construction grades command higher prices. Negative correlations were weak, with house_age and zipcode slightly lowering price.

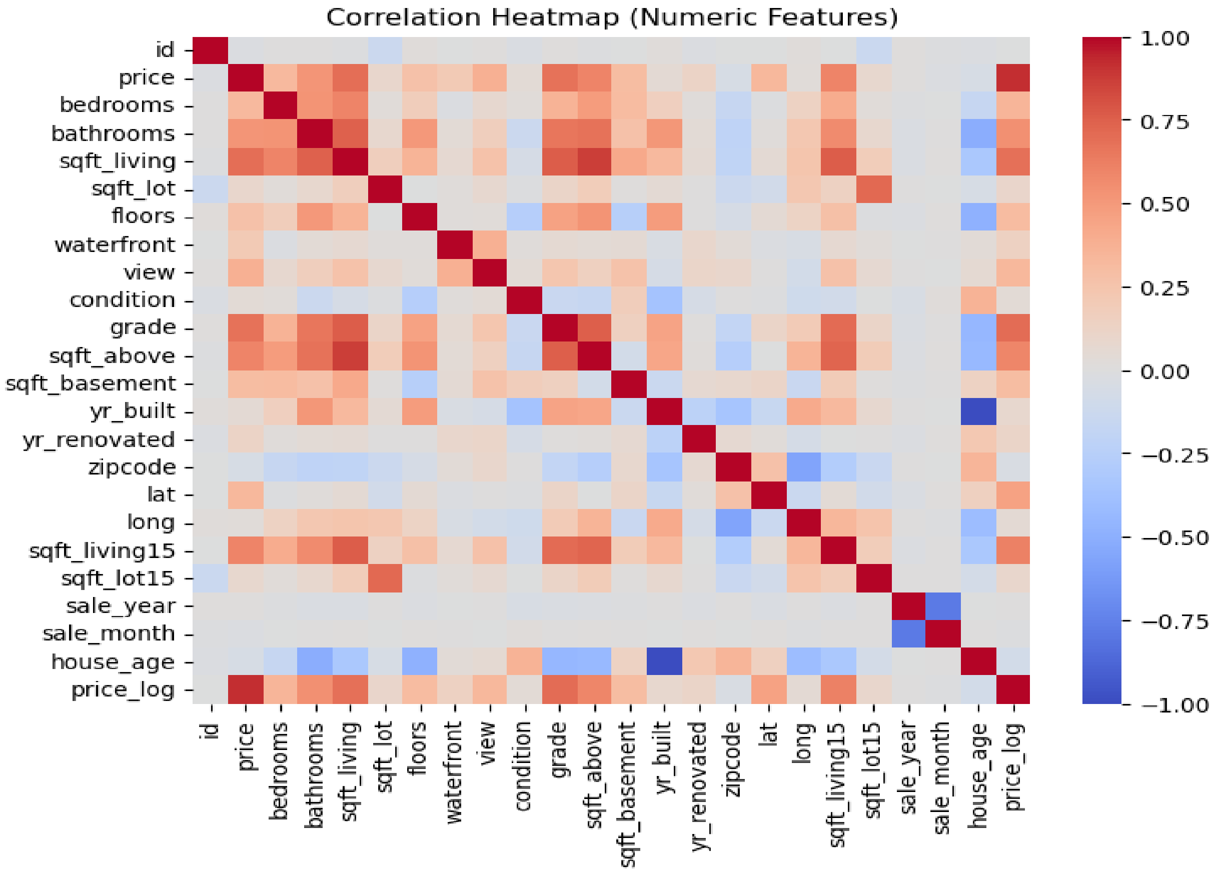


Figure 2: Correlation Heatmap (Numeric Features)

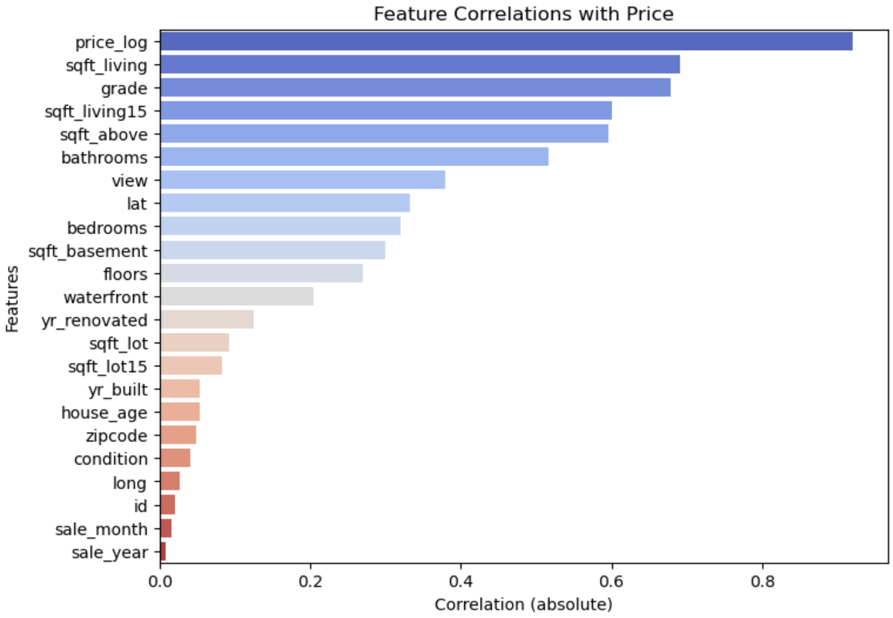


Figure 3: Feature Correlations with Price

The bar chart confirms that living area, construction grade, and lot size of nearby homes (sqft_living15) are top predictors of housing price.

c.Bivariate Relationships

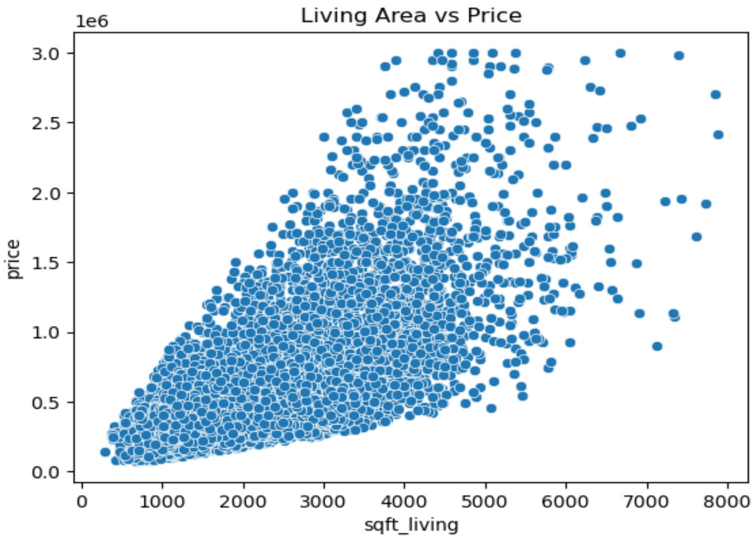


Figure 4: Living Area vs Price

Shows a clear positive trend as living space increases, home price rises, though at a diminishing rate beyond 4,000 sq. ft.

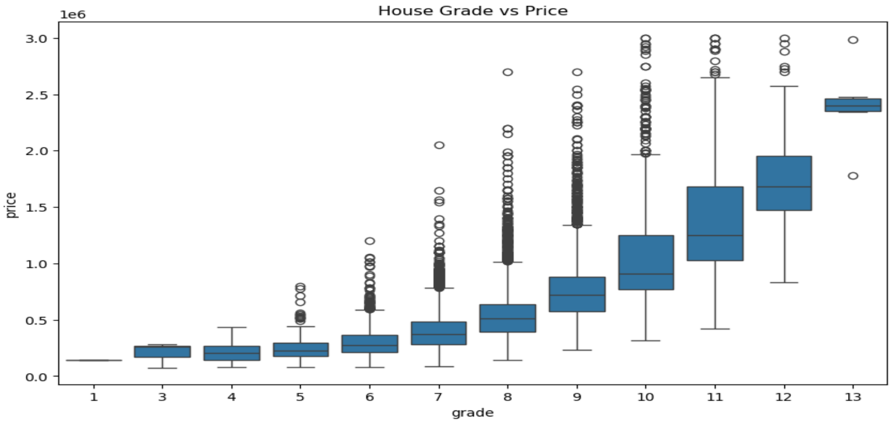


Figure 5: House Grade vs Price

Depicts how higher construction grades are directly associated with significantly higher property prices, highlighting grade as a strong qualitative predictor.

d. Summary of Findings

The exploratory analysis reveals that:

- Housing prices are primarily driven by living area, grade, and location (latitude/longitude).
- Bedrooms, bathrooms, and house age have smaller but still relevant effects.
- The dataset is clean, normally distributed after transformation, and suitable for building regression models.

Data Preparation for Modeling

After cleaning and feature engineering, the dataset was prepared for predictive modeling.

A subset of 13 explanatory variables was selected based on the strongest correlations with price and domain relevance.

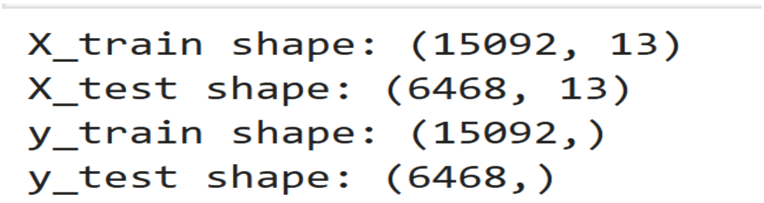
These include structural, quality, and location-based features:

- **Structural:** sqft_living, bedrooms, bathrooms, floors, house_age.
- **Quality:** grade, view, waterfront.
- **Location:** sqft_above, sqft_basement, lat, long, sqft_living15.

The data was divided into training (70%) and testing (30%) sets using the train_test_split() function.

This split ensures the models are evaluated on unseen data, preventing overfitting and enabling fair performance comparison.

Output Summary:



```
X_train shape: (15092, 13)
X_test shape: (6468, 13)
y_train shape: (15092,)
y_test shape: (6468,)
```

This confirms that the dataset was successfully partitioned, with 15,092 records for training and 6,468 for testing.

Experiments and Results

This section presents the experimental modeling phase of the analysis, where multiple supervised regression algorithms were applied to predict housing prices in King County. Each model was evaluated using the Root Mean Square Error (RMSE) and Coefficient of Determination (R2) metrics, as well as five-fold cross-validation to assess model consistency and generalizability. The selected models represent a progression from simple linear estimation to complex ensemble-based learning methods, demonstrating how increasing model complexity can improve predictive performance.

1.Linear Regression (Baseline Model)

Linear Regression served as the baseline model to estimate housing prices using 13 numerical and categorical predictors. This model assumes a linear relationship between the dependent variable (price) and the independent variables representing structural, quality, and location attributes.

After training on 70% of the data and testing on the remaining 30%, the model achieved the following results:

Model Performance:

- **RMSE:** 182,563.
- **R^2:** 0.711
- **Cross-Validation R^2:** 0.694 ± 0.008.

These values indicate that the model explains approximately 71% of the variance in house prices, which is a reasonable baseline given the dataset’s variability.

Additionally, a log-transformed version of the target variable was tested to reduce skewness and improve model fit. The log-based model yielded slightly improved results:

- **RMSE:** 176,934.
- **R^2:** 0.728.

The improvement in R^2 shows that addressing the right-skewed distribution of prices enhanced predictive accuracy slightly.

Feature Importance (Coefficients Analysis)

The figure below displays the relative impact of each predictor variable on the predicted price.

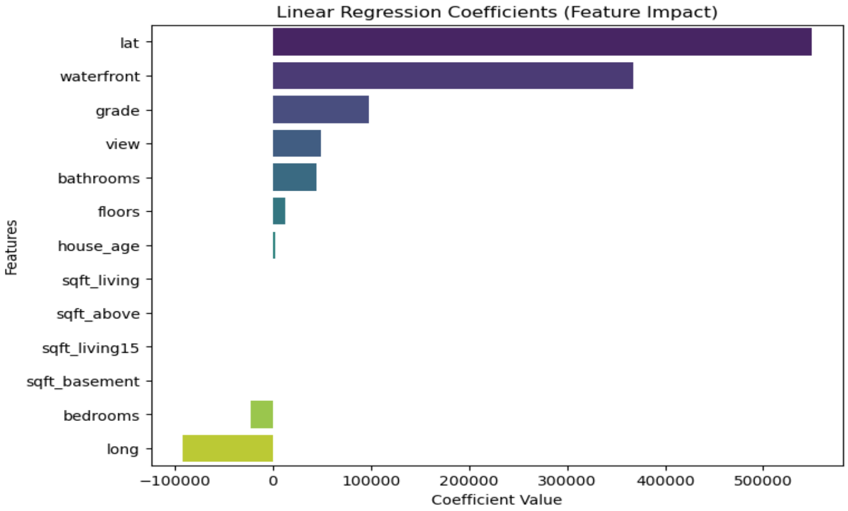


Figure 6: Linear Regression Coefficients (Feature Impact)

Bar chart showing influence of predictors such as latitude, grade, and waterfront.

Interpretation:

- Latitude (lat) and waterfront have the strongest positive effects on housing price, confirming that location and scenic views significantly increase home value.
- Grade, representing construction quality, is also major determinant, suggesting that better-quality homes sell at a premium.
- Bathrooms and view contribute moderately to price increases.
- Bedrooms and longitude (long) show negative coefficients, implying that homes farther east (lower longitudes) or with more but smaller bedrooms may have slightly reduced value, possibly reflecting smaller living space per room.

Table 2: Model Summary table

Metric	Linear Regression	Linear Regression (Log Target)
RMSE	182,563	176,934
R^2	0.711	0.728
Cross-Validation R^2	0.694± 0.008	-

2.Random Forest Regression (Ensemble Model)

To improve prediction accuracy and capture nonlinear relationships among housing features, a Random Forest Regressor was implemented. This ensemble methods builds multiple decision trees and averages their outputs, which reduces overfitting and increases predictive robustness.

Model Parameters:

- n_estimators = 300 (number of trees)
- max_depth = None (allow full tree growth)
- random_state = 42
- n_jobs = -1 (parallel computation for efficiency)

Model Performance

- **RMSE:** 124,225
- **R^2:** 0.866
- **Cross-Validation R^2:** 0.861± 0.010

Compared to Linear Regression (R^2 = 0.728), the Random Forest model provides a substantial improvement in predictive performance. The higher R^2 and lower RMSE indicate that the model explains about 86.6% of the variance in housing prices, suggesting that non-linear and interaction effects between variables are well captured.

Feature Importance

The figure below shows the relative contribution of each feature to the Random Forest model's prediction.

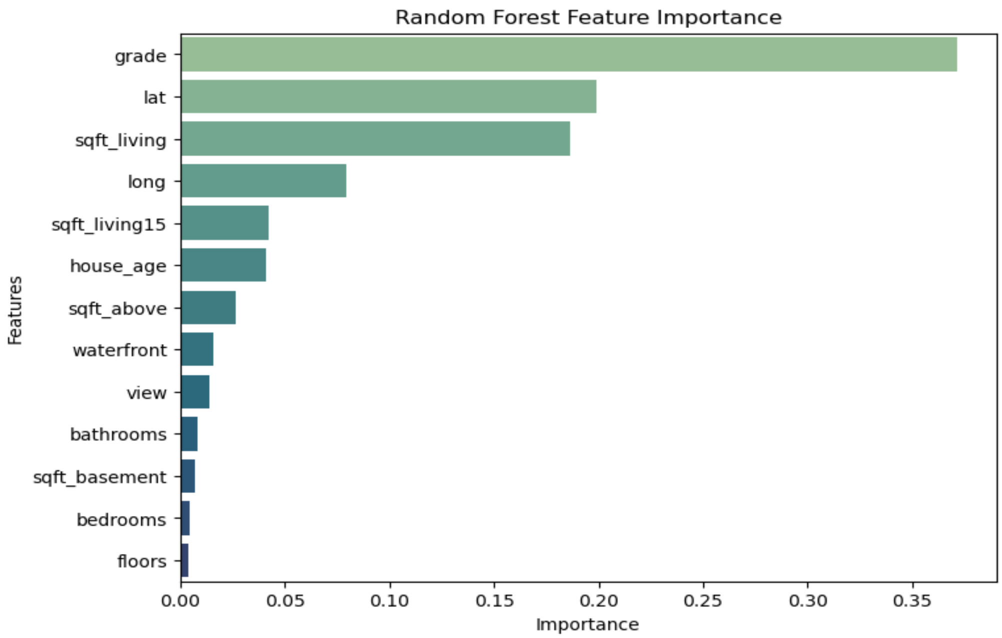


Figure 7: Random Forest Feature Importance

Bar chart illustrating feature contributions, showing grade, lat, and sqft_living as top driver of price.

Interpretation

- Grade is the most influential factor (importance = 0.37), reaffirming that better construction quality significantly drives up home prices.
- Latitude (lat) and square footage of living area (sqft_living) follow closely, highlighting that both geographic location and home size are key price determinants.
- Longitude (long), sqft_living15, and house_age have moderate importance, suggesting that neighborhood characteristics and property age contribute meaningfully.
- Bedrooms, floors, and bathrooms show relatively low importance, implying diminishing returns from additional rooms when size and quality are already considered.

Table 3: Model Summary Table

Metric	Linear Regression	Random Forest Regression
RMSE	176,934	124,224
R^2	0.738	0.866
Cross-Validation R^2	0.694± 0.008	0.861± 0.010

3.XGBoost Regression (Boosted Ensemble Model)

The third and most advanced approach used in this analysis is Extreme Gradient Boosting (XGBoost), an optimized gradient-boosted tree algorithm designed to improve both predictive power and computational efficiency. Unlike Random Forest, which averages multiple independent trees, XGBoost builds trees sequentially, where each new tree corrects the errors made by previous ones. This allows it to capture subtle patterns and nonlinear relationships more effectively.

Model Parameters:

n_estimators = 500.

learning_rate = 0.05

max_depth =6

subsample= 0.8

colsample_bytree = 0.8

random_state = 42

Model Performance

RMSE: 111,571

R^2: 0.892

Cross-Validation R^2: 0.883± 0.007

These results indicate that XGBoost achieved the best predictive performance among all three models. With an R^2 of approximately 0.89, the model explains nearly 90%of the variance in housing prices, a notable improvement over Random Forest ($R^2 = 0.866$) and Linear Regression ($R^2 = 0.711$). The lower RMSE further confirms that XGBoost provides the most accurate predictions overall.

Feature Importance

The chart below displays each feature's relative importance in determining home prices.

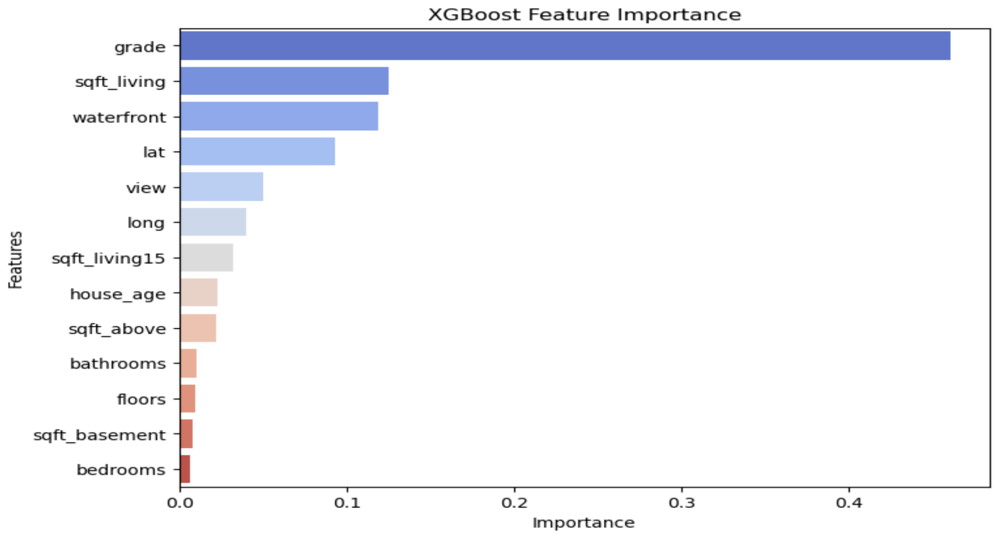


Figure 8: XGBoost Feature Importance

Interpretation

- Grade (quality of construction) remains the most dominant factor (importance = 0.46), reinforcing that higher-quality properties command significantly higher prices.
- Living area (sqft_living) and waterfront presence follow as major drivers, reflecting the premium associated with spacious homes and scenic locations.
- Latitude (lat) and view also influence prices, emphasizing geographic desirability and aesthetic appeal.
- Less impactful features include bedrooms, floors, and sqft_basement, which contribute marginally once major attributes like quality and size are accounted for.

4.Ridge Regression (Regularized Linear Model)

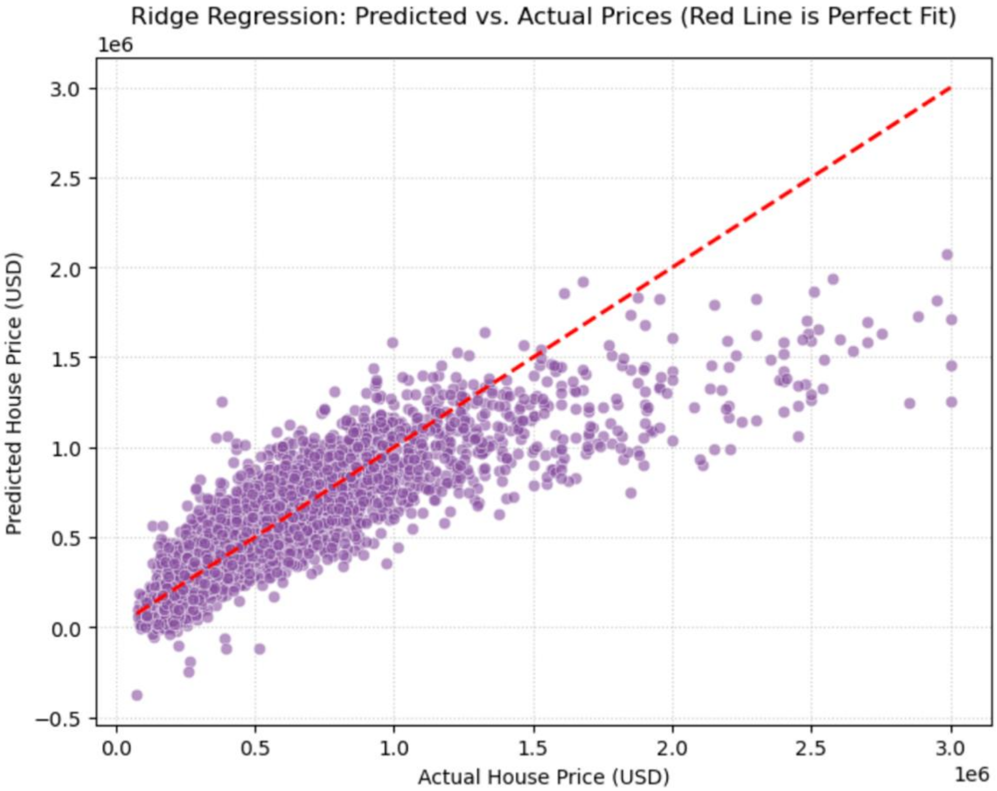
Ridge Regression is a linear modeling technique that introduces **L2 regularization** to minimize overfitting and stabilize coefficient estimates. It is especially effective when predictors are highly correlated or when the model tends to overfit due to a large number of features. Before training, all numerical features were standardized using **StandardScaler** to ensure that each feature contributes equally to the model's penalty term.

Model Parameters

- Regularization strength (α): 10.0
- Solver: Auto (default)
- N B bRandom state: 42

Feature Importance

This graph below illustrates the relationship between predicted and actual home prices. The red dashed line represents the ideal 45° fit line (perfect predictions). Points scattered closer to this line indicate stronger model accuracy.



Interpretation

- Ridge regularization effectively **reduces overfitting** compared to Ordinary Least Squares regression by constraining coefficient magnitudes.
- However, it still assumes a **linear relationship** between predictors and target values, which limits its performance on complex real estate data.
- Despite these limitations, Ridge Regression provides valuable interpretability, highlighting the relative linear influence of major features such as living area, grade, and latitude.
- It serves as a **baseline regularized model** against which more advanced methods (like Gradient Boosting or XGBoost) can be compared.

Table4: Model Summary Table

Metric	Test Set	Cross-Validation (Mean ± Std)
R² Score	0.7107	0.6935 ± 0.0077
RMSE (USD)	182,568.54	—

These results show that the Ridge Regression model explains approximately **71% of the variance** in house prices on the test set. While it performs better than the simple Linear Regression baseline, it lags behind more complex ensemble models like Random Forest and XGBoost. The relatively higher RMSE also suggests that the linear assumption limits the model's ability to capture nonlinear patterns in housing data.

Model Comparison

(a) Model Performance (R² Comparison)

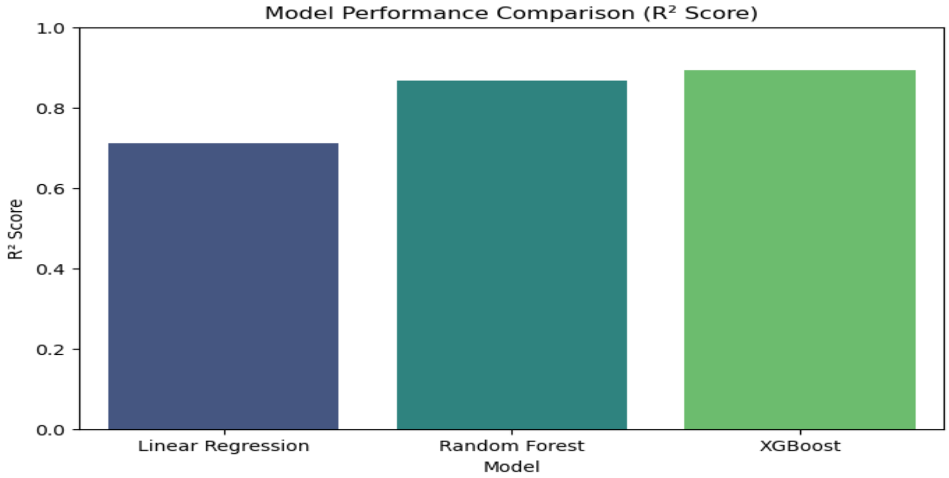


Figure 9: Model Error Comparison (R² Score)

The above figure shows the R² comparison across models. XGBoost achieved the highest R² of 0.88, followed by Random Forest (0.86) and Linear Regression (0.71). This indicates that ensemble methods capture non-linear relationships better, leading to improved prediction accuracy.

(b) Model Error (RMSE Comparison)

The below figure shows model RMSE values. XGBoost had the lowest RMSE ($112,000), followed by Random Forest ($124,000), and Linear Regression ($182,000). Lower RMSE indicates that XGBoost provides the most precise price estimates.

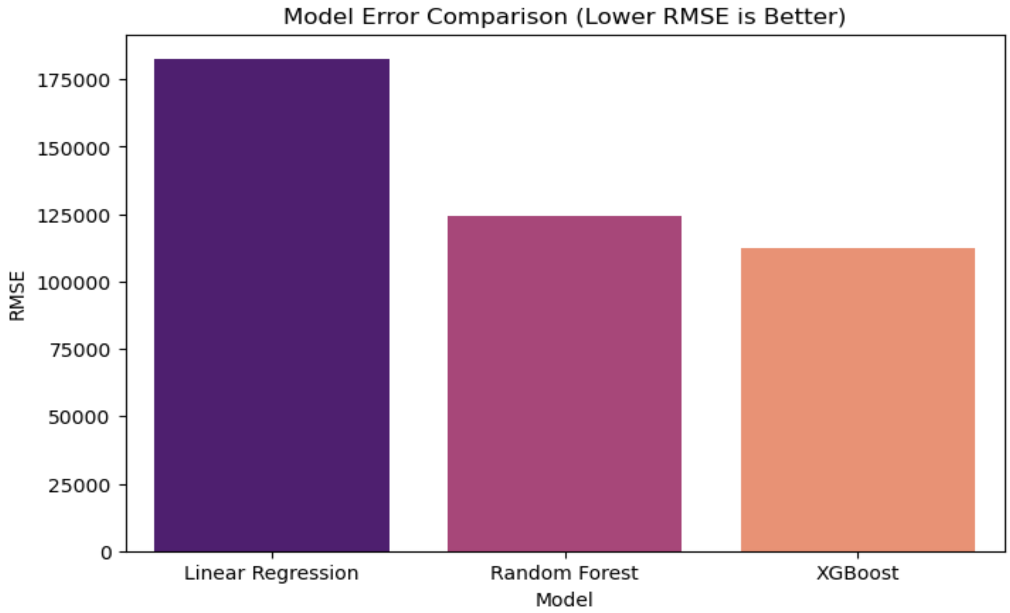


Figure 10: Model Error Comparison (Lower RMSE is Better)

Table 4: Model Summary Table

Metric	**Linear Regression**	**Random Forest**	**XGBoost**
RMSE	176,934	124,225	111,571
R²	0.728	0.866	0.892
Cross-Validation R²	0.694 ± 0.008	0.861 ± 0.010	0.884 ± 0.006

Conclusion

This project presents a comprehensive and data-driven investigation of determining factors for housing prices in King County, combining statistical rigor with predictive modeling sophistication. The project lays a strong foundation through systematic data preprocessing, feature engineering, and model evaluation to extract a simple relationship between home characteristics, grades of construction quality, and geographic desirability as they determine prices for homes. The model ultimately tells us that living area, quality of construction, and latitude/longitude are the influential predictors to real-estate value that suggest physical and locational reasons for capturing value.

Amongst the models that were evaluated, XGBoost Regression is the most superior at predicting values, with a R² value of 0.892, and lowest RMSE (~$111,571). It outperformed both Random Forest and Linear Regression substantially, and indicates the strength associated with boosted ensemble models to model complex and non-linear data structures to identify subtle features mutually accept attribution outside of the traditional norm. Lastly, a linear model, like Ridge Regression, gives valuable interpretability, however its assumptions simplify the predicting results when applied to heterogeneous and miscellaneous datasets, such as real estate..

The consequences of this research reach beyond mere academic modeling; the conclusions can be of use to real-estate analysts, investors, and urban policymakers in regards to property valuation, investment holdings, and housing policy. The analysis structure developed here—integrating data cleansing, EDA, and machine learning—can serve as an example and model for others looking to implement predictive analytics in this realm. In advocating transparency, rigorous methodology, and validation, this capstone supports the crucial role of machine learning in bolstering properties' undertone capacity as data—and in evolving that data into information in the marketplace—allowing for data-driven real-estate forecasting.

Contributions

This capstone project was completed collaboratively by **Capstone Group 03**, with each member contributing equally to different phases of the analysis.

Meghana Yalam identified and sourced the King County House Sales dataset from Kaggle and developed the initial research questions and report. She also performed analysis and model implementation for the **Ridge Regression (Regularized Linear Model** method.

Prasanthi Chitturi conducted the **Exploratory Data Analysis (EDA)**, including data cleaning, feature engineering, and visualization of key relationships in the dataset. She also implemented and evaluated the **Linear Regression, Random Forest** model and contributed to compiling the written report.

Tharun Pallela assisted with documentation and presentation preparation, and developed the **XGBoost** model. He also reviewed the report for formatting, clarity, and submission requirements.

All members collaborated to interpret the results, review visual outputs, and ensure the final report met the technical and academic standards of the course.

References

1.Kaggle. (2016). *House Sales in King County, USA* [Data set]. Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

2.Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

3.Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

4.Breiman, L. (2001). *Random forests.* Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

5.Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

6.Waskom, M. L. (2021). *Seaborn: Statistical data visualization.* Journal of OpenSource Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

7.McKinney, W. (2010). *Data structures for statistical computing in Python.* In *Proceedings of the 9th Python in Science Conference* (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

8.King County Department of Assessments. (2015). *Property sales data for King County, Washington (2014–2015).* Public Records, King County, WA.