Humber College Institute of Technology & Advanced Learning

Estate Estimator: Machine Learning for Competitive Property Pricing

Jyothi Swaroop Muralasetti

Abstract :

This study explores machine learning techniques for house price prediction using data from King County, Washington. We implemented Linear Regression, K-Nearest Neighbors, and Random Forest algorithms after applying feature selection and PCA for dimensionality reduction. Property size, quality, and location emerged as key price determinants, with Random Forest using all features achieving the best performance (RMSE: $117,528.94). Various machine learning algorithms are discussed, highlighting their applications in data mining, image processing, and predictive analytics, with the key advantage being their ability to work automatically once trained. The real estate market is a complex ecosystem influenced by numerous factors ranging from property characteristics to location dynamics. Accurately predicting house prices is crucial for various stakeholders including buyers, sellers, real estate agents, and financial institutions. Through careful pre-processing, feature engineering, and model development, we aimed to build a robust prediction system that can accurately estimate house prices based on property attributes. The findings demonstrate machine learning's effectiveness in real estate valuation with practical applications for market stakeholders.

1. Introduction

The real estate market is a complex ecosystem influenced by numerous factors ranging from property characteristics to location dynamics. Accurately predicting house prices is crucial for various stakeholders including buyers, sellers, real estate agents, and financial institutions. This report presents a comprehensive analysis of house price prediction using various machine learning techniques, focusing on feature selection, dimensionality reduction, and model performance comparison.

Our analysis utilizes a housing dataset containing information about properties in King County, Washington. Through careful pre-processing, feature engineering, and model development, we aim to build a robust prediction system that can accurately estimate house prices based on property attributes. This project not only demonstrates the application of machine learning in real estate valuation but also provides insights into the key factors driving property prices in the market.

2. Data Overview and Pre-processing

2.1 Dataset Description

The dataset contains 21,613 records of house sales with 21 attributes for each property. These attributes include:

- Basic property information: bedrooms, bathrooms, square footage (living space, lot, basement)
- Location data: latitude, longitude, zip code
- Property condition and quality: grade, condition, view rating
- Age information: year built; year renovated
- Other features: floors, waterfront property, etc.

The target variable is the sale price of the property, which ranges from $75,000 to several million dollars.

2.2 Data Cleaning Process

Several pre-processing steps were implemented to prepare the data for analysis:

1. **Data Type Conversion**: Converting appropriate columns to their correct data types:
   - 'date' converted to date-time format
   - 'zipcode' converted to string type
   - 'lat' and 'long' ensured as float type
2. **Handling Missing Values**: The data contained minimal missing values, with only 2 missing entries in the 'sqft_above' column. These were imputed with the median value of the column to maintain data integrity.
3. **Outlier Detection and Removal**: Price distribution analysis revealed significant outliers at the upper end of the price spectrum. Properties priced above $2.5 million (97 records) were removed as outliers to create a more representative model, bringing our working dataset to 21,516 records.
4. **Feature Selection**: After correlation analysis, three columns ('id', 'date', and 'zipcode') were removed as they had low correlation with the target variable and would not contribute meaningfully to prediction accuracy.

Table 1: Summary Statistics After Pre-processing

Statistic	*Price ($)*	*Bedrooms*	*Bathrooms*	*Sqft_living*	*Grade*
Count	21,516	21,516	21,516	21,516	21,516
Mean	534,883	3.37	2.11	2,080	7.66
Min	75,000	0	0	290	1
25%	321,950	3	1.75	1,427	7
50%	450,000	3	2.25	1,910	7
75%	645,000	4	2.5	2,550	8
Max	2,500,000	33	8	13,540	13

Statistics of the Data:

- The dataset contains 21,516 housing records with comprehensive information about prices and key property characteristics. The average house price is $534,883, with values ranging widely from $75,000 to $2,500,000. The middle 50% of properties fall between $321,950 and $645,000, with a median price of $450,000.
- Regarding bedrooms, properties average 3.37 bedrooms, with a median of 3. While some properties have no bedrooms (minimum of 0), there's an outlier with 33 bedrooms. Most homes (middle 50%) have between 3-4 bedrooms.
- For bathrooms, the average is 2.11, with a median of 2.25. The range extends from 0 to 8 bathrooms, with the middle 50% of properties having between 1.75 and 2.5 bathrooms.
- Living space (sqft_living) averages 2,080 square feet, with a median of 1,910 square feet. This feature shows considerable variation, ranging from a compact 290 square feet to a spacious 13,540 square feet. The middle 50% of properties have between 1,427 and 2,550 square feet.
- The grade measure, which likely represents construction quality, averages 7.66 with a median of 7. The values range from 1 to 13, with the middle 50% of properties rated between 7 and 8, suggesting most homes are of average to above-average quality.

3. Exploratory Data Analysis

3.1 Price Distribution

A thorough analysis of the price distribution revealed a right-skewed pattern typical of real estate pricing data. Most properties in the dataset are priced between $300,000 and $650,000, with the frequency decreasing as price increases. After removing outliers above $2.5 million, the price distribution became more manageable for modelling purposes.

3.2 Correlation Analysis

A correlation matrix was generated to identify the key features that influence house prices. The top 6 features most strongly correlated with price were:

1. Square footage of living space (sqft_living): 0.70
2. Grade of the house (grade): 0.67
3. Square footage above ground (sqft_above): 0.61
4. Square footage of living space in 2015 (sqft_living15): 0.59
5. Number of bathrooms (bathrooms): 0.53
6. View quality (view): 0.40

This analysis provided valuable insights into which features would likely be most predictive in our models. Interestingly, while the number of bedrooms shows a positive correlation with price (0.31), it's not as strong as one might expect, suggesting that the quality of space (as measured by square footage and grade) is more important than the raw number of rooms.

3.3 Feature Relationships

Several relationships between features were observed:

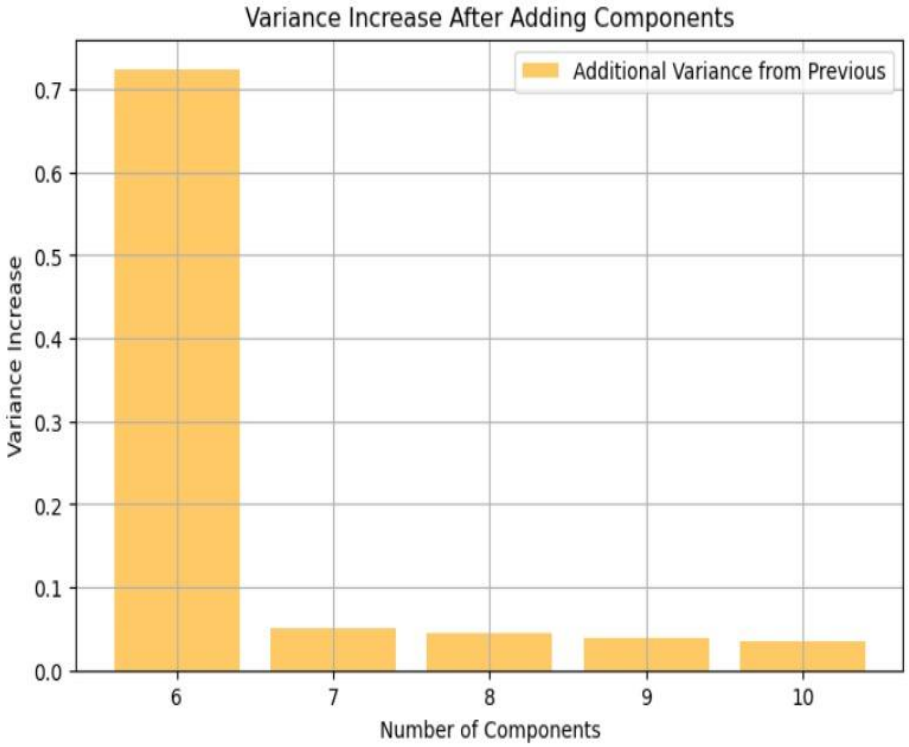
- Strong correlation between sqft_living and bathrooms (0.75)
- Moderate correlation between sqft_living and bedrooms (0.58)
- Strong correlation between grade and sqft_living (0.76)
  These relationships indicate some multicollinearity in the data, which was later addressed through dimensionality reduction techniques.

4. Feature Engineering and Selection

4.1 Principal Component Analysis (PCA)

To address multicollinearity and reduce dimensionality, Principal Component Analysis was applied to the standardized feature set. A cumulative variance analysis was conducted to determine the optimal number of components:

Fig 1: PCA Cumulative Variance Analysis



Based on the analysis, 6 principal components were selected as they captured approximately 72% of the variance in the data. The rate of additional variance explained dropped significantly after 6 components, suggesting diminishing returns for including more components.

4.2 Backward Elimination

As an alternative to PCA, we implemented a backward elimination technique to select the most significant features based on statistical significance. This process iteratively removed the least significant features until only the most impactful ones remained. The top 6 features selected through backward elimination were:

1. View quality (view)
2. Grade (grade)
3. Square footage above ground (sqft_above)
4. Square footage of basement (sqft_basement)
5. Year built (yr_built)
6. Latitude (lat)

This feature selection method provided a complementary approach to PCA, focusing on interpretability rather than mathematical transformation.

5. Model Development and Evaluation

We implemented and compared several machine learning algorithms for house price prediction. Each model was trained and evaluated using both the PCA-transformed data and the feature-selected data.

5.1 Linear Regression

Linear regression models were trained on both PCA components and top selected features:

- Linear Regression with PCA: RMSE = $188,981.20
- Linear Regression with Top Features: RMSE = $193,070.49

5.2 K-Nearest Neighbours Regression

KNN regression was implemented with different k values (5 and 10) on the PCA-transformed data:

- KNN-5 with PCA: RMSE = $153,027.17, MAE = $91,540.66
- KNN-10 with PCA: RMSE = $155,071.66, MAE = $91,515.45

5.3 KNN Classification

5.3 KNN Classification

For comparison, we also implemented a classification approach by converting the continuous price variable into three distinct categories:

This transformation divided house prices into three equally-sized quantiles (low, medium, and high price ranges) labelled as 0, 1, and 2. KNN classification was then applied with two different values of k:

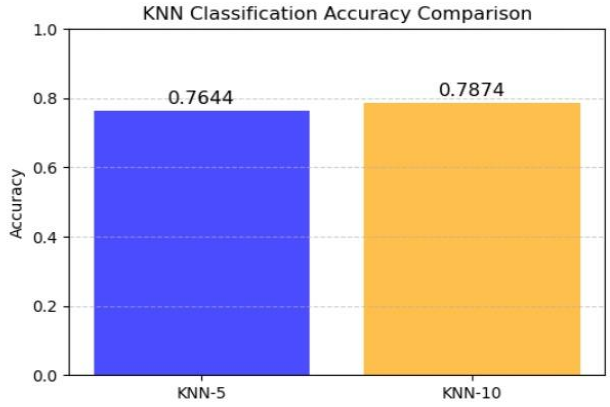


Fig 2: Comparison of Accuracies of KNN_5 and KNN_10

KNN-5 Classification Accuracy: 76.44%

KNN-10 Classification Accuracy: 78.74%

The higher accuracy for KNN-10 suggests that using more neighbours provides better classification performance in this context. This classification approach offers an alternative perspective, particularly useful for scenarios where estimating the price range of a property is more important than predicting its exact price.

5.4 Random Forest Regression

Random Forest models were implemented with various feature sets:

- Random Forest with All Features: RMSE = $117,528.94
- Random Forest with Top Features: RMSE = $153,280.21
- Random Forest with PCA: RMSE = $150,520.18

5.5 Model Performance Comparison

Table 2: Model Performance Comparison (RMSE)

Model	***RMSE***
Random Forest (All Features)	$117,528.94
Random Forest (PCA)	$150,520.18
KNN-5 (PCA)	$153,027.17
Random Forest (Top Features)	$153,280.21
KNN-10 (PCA)	$155,071.66
Linear Regression (PCA)	$188,981.20
Linear Regression (Top Features)	$193,070.49

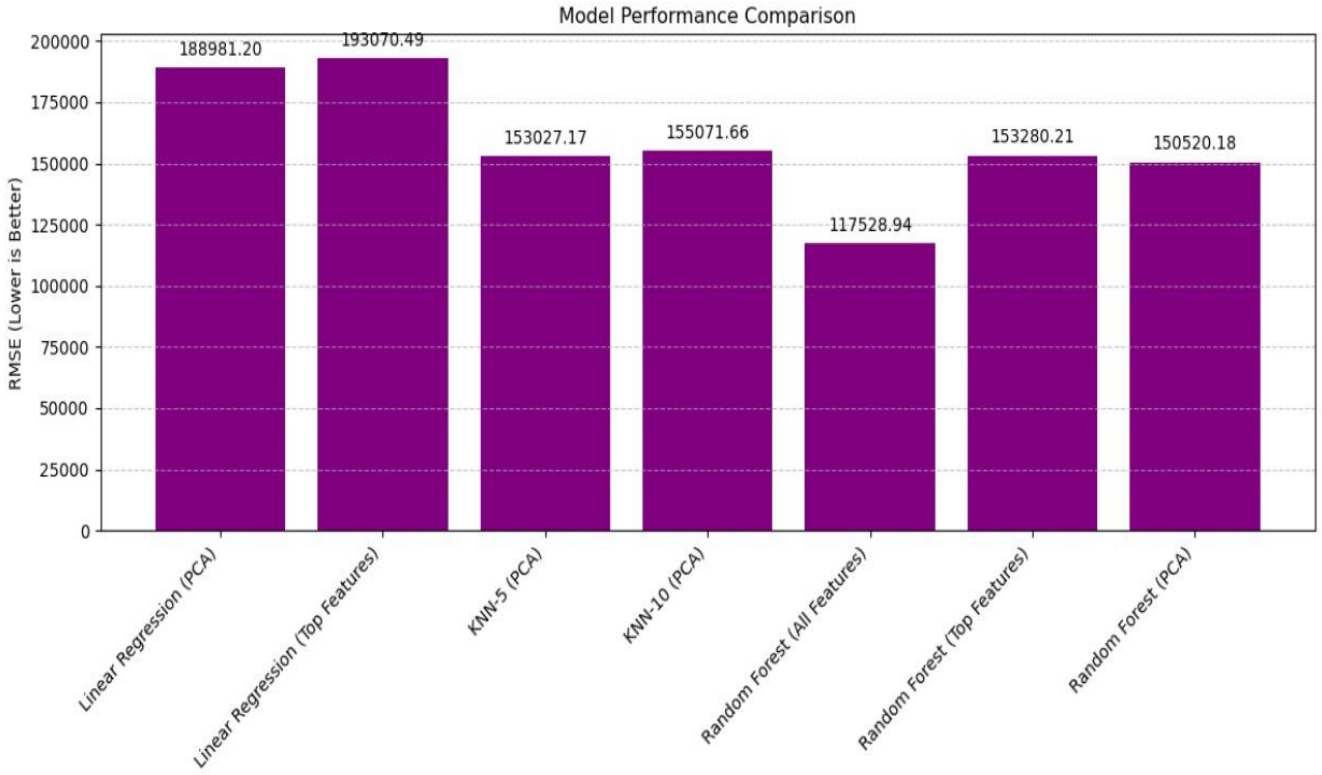


Fig 4: Comparison of RMSE Values of all models used

The Random Forest model with all features performed the best with the lowest RMSE of 117,528, making it the most accurate predictor. KNN models performed moderately well, while Linear Regression had the highest errors, especially with PCA. Feature selection and PCA helped reduce dimensionality but impacted accuracy slightly. Future work includes optimizing hyperparameters, exploring ensemble models, and incorporating location-based features for better predictions.

6. Case Study: Sample House Price Predictions

To illustrate the practical application of our models, we predicted prices for two sample houses from the test set using different methods.

Table 3: House 1 Price Predictions

Method	***Price ($)***
Actual Price	245,000.00
Predicted (Linear Regression)	234,711.00
Predicted (KNN-5)	275,640.00
Predicted (KNN-10)	293,919.00
Predicted (RF-All Features)	251,784.00
Predicted (RF-Top Features)	279,570.00
Predicted (RF-PCA)	259,131.00

Table 4: House 2 Price Predictions

Method	***Price ($)***
Actual Price	641,000.00
Predicted (Linear Regression)	518,931.00
Predicted (KNN-5)	629,400.00
Predicted (KNN-10)	616,391.50
Predicted (RF-All Features)	634,408.00
Predicted (RF-Top Features)	574,144.00
Predicted (RF-PCA)	610,208.00

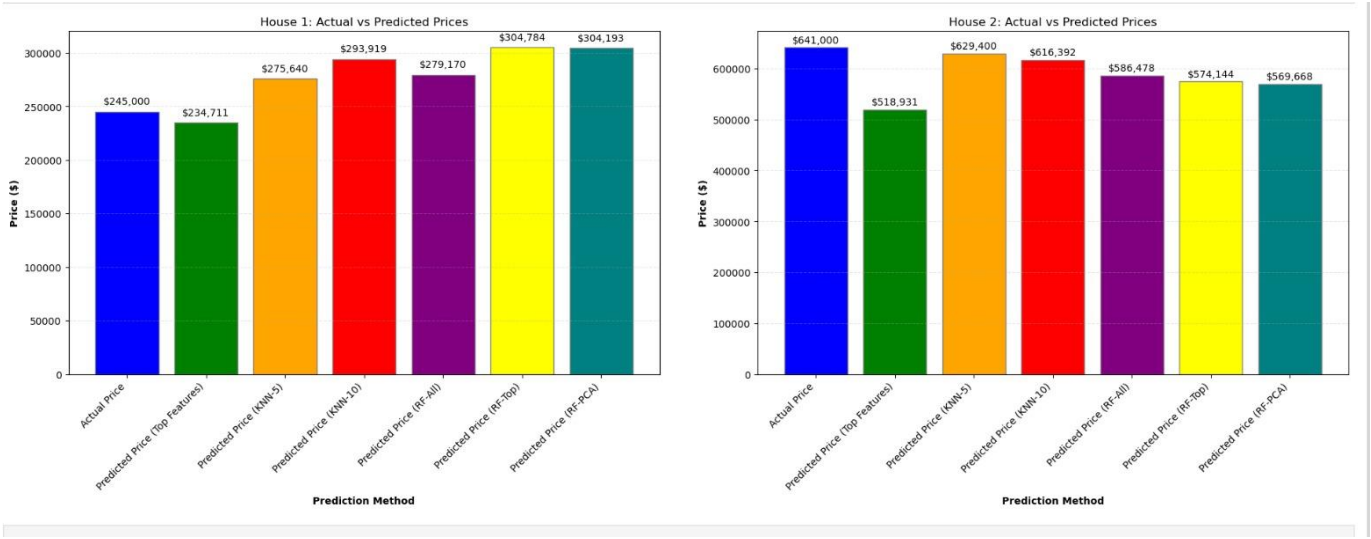


Fig 5 : Comparison of Actual vs Predicted Prices using all models used used

The case study reveals that the Random Forest model with all features provides the closest predictions to the actual prices, particularly for House 1. For House 2, KNN-5 provided the closest prediction, though Random Forest with all features was also very accurate.

7. Discussion

7.1 Feature Importance

The analysis revealed several key insights about house price determinants:

1. **Size Matters**: Square footage of living space consistently emerged as the strongest predictor of house price, confirming the intuitive relationship between property size and value.
2. **Quality over Quantity**: The grade of the house (which rates the construction quality) proved more important than the number of bedrooms, suggesting that quality trumps quantity in the real estate market.
3. **Location Significance**: The latitude variable's importance in the backward elimination process underscores the real estate mantra of "location, location, location."

7.2 Model Comparison

Our findings demonstrate that:

1. **Random Forest Superiority**: The Random Forest algorithm outperformed other models significantly, likely due to its ability to capture complex non-linear relationships and handle the interaction effects between features.
2. **Feature Selection vs. All Features**: Interestingly, using all features with Random Forest yielded better results than using selected features, suggesting that despite potential multicollinearity, the algorithm benefits from the additional information contained in seemingly less important features.
3. **PCA Effectiveness**: While PCA didn't lead to the best model, it did improve the performance of Linear Regression and provided competitive results with relatively few components, demonstrating its value for dimensionality reduction.

7.3 Limitations

Despite the promising results, our analysis has several limitations:

1. **Geographical Constraints**: The dataset is limited to King County, Washington, which may limit the generalizability of findings to other real estate markets.
2. **Temporal Factors**: The dataset covers a limited time period, and real estate markets are known to fluctuate over time. Models might need periodic retraining to remain accurate.
3. **Missing External Factors**: Important external variables like interest rates, economic indicators, and neighborhood development plans were not included in the analysis.
4. **Computational Limitations for Feature Selection**: Attempts to implement stepwise elimination for feature selection were hindered by insufficient processing power on the available hardware. This computational

constraint prevented a more thorough exploration of optimal feature combinations, potentially leaving valuable predictive patterns undiscovered or resulting in a less-than-optimal feature subset in the final models.

8. Conclusions

This comprehensive analysis of house price prediction using machine learning techniques has yielded several important conclusions:

1. The Random Forest algorithm using all available features provides the most accurate price predictions, with an RMSE of $117,528.94.
2. The most influential factors in determining house prices are the square footage of living space, the quality grade of the house, and the square footage above ground.
3. While dimensionality reduction through PCA and feature selection through backward elimination didn't outperform using all features, they provided valuable insights and still produced reasonably accurate models.
4. For practical applications, the choice of model might depend on the specific requirements: Random Forest offers the highest accuracy but at the cost of interpretability, while Linear Regression provides more transparent coefficients that can be directly interpreted.
5. The success of machine learning in this domain demonstrates its potential for revolutionizing real estate valuation, potentially benefiting all stakeholders in the market.

9. References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer Science & Business Media.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.