**“Machine Learning: The Key to Smoother Real Estate Transactions”**  
  


**Real Estate Business Consultants & Managers:**

|  |  |
| --- | --- |
| **Ayushi Arora** | **Krupali Patel** |
| **Albira Nousin** | **Tashmeet Kaur Saluja** |

**Table of Contents**

1. Introduction
2. Data Overview
3. Literature Review
4. Methodology
5. Model Selection and Results
6. Comparison of Prediction Models
7. Business Implications
8. Key takeaways from model comparison
9. Conclusion
10. Reference

**1. Introduction**

The real estate market plays a pivotal role in the economy. Accurate property price predictions enable businesses, investors, and homebuyers to make well-informed decisions. Predicting property prices, influenced by factors like location, square footage, and amenities, offers valuable insights that can drive business strategies. This report explores how machine learning algorithms, such as Linear Regression (LR), K-Nearest Neighbors (KNN), Random Forest, and XGBoost, can help predict house prices based on historical data, with the goal of improving pricing, investment, and strategic decisions in the real estate sector.

**2. Data Overview**

The kc\_house\_data dataset, consisting of 21,613 records and 21 features, forms the basis of our analysis. Key features include house prices, number of bedrooms, square footage, and location.

Key Statistics from the Data:

* Number of Records: 21,613
* Number of Features: 21 (e.g., bedrooms, bathrooms, square footage)
* Target Variable: Price (house price)

**Business Relevance:**

* Understanding which features (like square footage, location, and amenities) correlate most strongly with house prices allows real estate businesses to refine their pricing strategies, determine the best investment opportunities, and tailor marketing efforts to target key demographics.

**3. Literature Review**

Machine learning has become a powerful tool in predicting house prices, offering a more accurate and efficient alternative to traditional methods. Key machine learning models like Linear Regression and KNN can predict prices by analyzing historical patterns and local trends.

* Linear Regression: Works well for simple relationships between features and prices.
* K-Nearest Neighbors (KNN): Can model more complex, non-linear relationships by comparing properties to nearby ones.
* Challenges: Real estate markets are volatile and influenced by external factors (economic shifts, unexpected events) that are often hard to predict using historical data alone.

**4. Methodology**

To predict house prices, we performed feature selection, model training, and evaluation:

We have made ‘view’ out dummy variables as the other categorical variables either would not need to be dummy or would have made the method of exploring house pricing more complicated. A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. (Trochim, n.d.)

Feature Selection: We selected six features strongly correlated with the target variable, price:

* sqft\_living: Living space in square feet (Directly correlates with price)
* grade: Overall quality of the house (Directly influences the price)
* sqft\_above: Above-ground living space (Key determinant of price)
* sqft\_living15: Size of the house 15 years ago (Helps capture long-term trends)
* bathrooms: Number of bathrooms (Affects both price and size)
* sqft\_basement: Basement size (Can significantly add value to the property)

**Figure 1: Correlation percentages with price in (desc) order**

**5. Model Selection and Results**

We compared four different machine learning models based on their performance in predicting house prices. Below are the predicted house prices for two sample houses:

**1. Linear Regression (LR)**

* House 1 Price: $320,020.56
* House 2 Price: $545,292.41

R² value: 0.5481 (Slightly more accurate than KNN)

**2. K-Nearest Neighbors (KNN) (K=10)**

* House 1 Price: $332,290.00
* House 2 Price: $620,284.60

R² value: 0.5418 (Shows a good fit, but slightly less accurate than LR)

**3. Random Forest**

* House 1 Price: $256,332.00
* House 2 Price: $606,687.92

R² value: 0.5860 (More stable than KNN)

**4. XGBoost**

* House 1 Price: $339,528.81
* House 2 Price: $614,605.31

R² value: 0.6025 (Shows high accuracy but greater variability)

**6. Comparison of Prediction Models**

* Linear Regression: Offers simplicity and interpretability with an R² of 0.5481, making it slightly more accurate than KNN.
* KNN: More flexible for capturing local patterns but has a slightly lower R² value of 0.5418.
* Random Forest: Offers better accuracy with R² of 0.5860 and provides stability by averaging multiple decision trees.
* XGBoost: Most efficient and accurate, achieving R² of 0.6025, but has some instability in predictions.

**Table 1: House prices after different model implementation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | R² | House\_1\_Predicted\_Price | House\_2\_Predicted\_Price |
| Linear Regression | $164,638.78 | $68,398,575,519.50 | $261,531.21 | 0.5481 | $320,020.56 | $545,292.41 |
| KNN (K=10) | $168,295.14 | $69,343,757,263.47 | $263,332.03 | 0.5418 | $332,290.00 | $620,284.60 |
| Random Forest | $157,366.30 | $68,336,021,727.36 | $261,411.59 | 0.5485 | $256,332.00 | $606,687.92 |
| XGBoost | $155,421.28 | $77,177,430,621.23 | $277,808.26 | 0.4901 | $339,528.81 | $614,605.31 |

**7. Business Implications**

Predicting house prices accurately has substantial business implications:

* Investment Strategies: Knowing the most important features that influence price (like square footage and location) helps businesses make better investment choices.
* Pricing Decisions: More accurate predictions assist businesses in setting the right prices for homes, improving profitability and customer satisfaction.
* Marketing Targeting: With insights into the features most appealing to buyers, real estate agents can tailor marketing strategies more effectively.
* Beat competitors: The market is highly competitive, and housing is very expensive investment. Price is a huge factor here.

**8. Key Takeaways from Model Comparison**

* **Random Forest** achieves the best scores in three of the four metrics (MSE, RMSE, R²), indicating overall better performance in predictive accuracy.
* **XGBoost** has the lowest MAE, which may be advantageous for minimizing average prediction errors, but its higher RMSE and lower R² suggest it struggles with larger variances.
* **Conclusion**: **Random Forest** is the preferred model based on its balanced performance across most metrics, making it reliable and accurate for predicting house prices.

**9. Conclusion**

Machine learning models offer powerful tools for predicting house prices. While simpler models like Linear Regression provide interpretability, more complex models like Random Forest and XGBoost show superior accuracy, capturing the intricate relationships between features. Businesses in the real estate sector can leverage these models to refine pricing strategies, improve investment decisions, and better serve customers.

Real estate companies can use these models to stay competitive by predicting house prices more accurately, guiding both investors and homebuyers in their decision-making processes.  
  
  
**Reference:**

Trochim, W. M. (n.d.). Dummy variables. Research Methods Knowledge Base. https://conjointly.com/kb/dummy-variables/