CNST 6308 Data Analysis in Construction Management

Predicting House Prices Using Machine Learning and Artificial Neural Networks

Submitted by

Madhukumar Gopal (2354612)

Under the guidance of
Dr. Lu Gao

Contents

1. Abstract

The project uses machine learning and artificial neural networks to predict house prices at the core of construction management and urban development. In this study, house price prediction was done using the King County Housing Dataset [1] with appropriate methodologies like data preprocessing, feature importance analysis, and ensemble modeling. These are important attributes of property value and help in construction project planning, resource allocation, and market assessment. Integration of ensemble techniques, such as Random Forest and XGBoost, into ANN models further enhances its value for data-driven approaches toward improvement in decision making within construction management.

2. Introduction

Accurate house price prediction is of great importance to stakeholders in the business of real estate development, urban planning, and construction. Better decisions can be made under the processes involved in project planning, budgeting, and resource allocation. In a time when large scale real estate data is available, machine learning techniques have become a powerful tool in deriving insights to better predictive accuracy.

This project is using the King County Housing Dataset [1], which is large and has various information about house sales, including square footage, bedrooms, bathrooms, and many geospatial features. The focus here will be to develop a robust predictive model for house prices using leading edge machine learning algorithms in combination with artificial neural networks (ANNs).

Key elements of the project include:

Data Preprocessing: Handling missing values, scaling numerical features, and encoding categorical data to be compatible with machine learning algorithms.

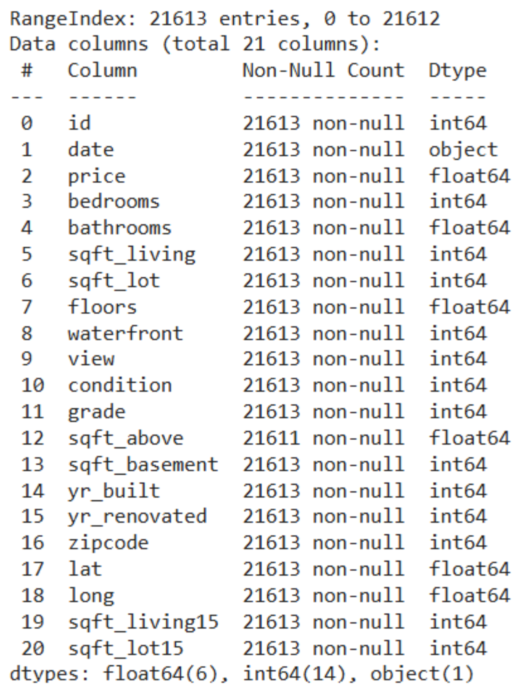
Feature Importance Analysis: Investigate crucial features of house price datasets using techniques like XGBoost and Random Forest.

Model Development: Comparing different models, including traditional regression, ensemble methods, and ANN architectures, to identify which approach would work best on house price prediction.

3. Dataset Description

The data set used in this study is the King County Housing Dataset[1], which contains data on house sales in King County, USA. It includes 21,613 rows and 21 features, covering various aspects of houses and their sale prices. (Fig.1)

- **Price**: The sale price of the house (e.g., $538,000).
- **Bedrooms and Bathrooms**: The number of bedrooms and bathrooms in the house.
- **Living Space**: The total size of the indoor living area, measured in square feet.
- **Lot Size**: The total size of the property, including outdoor areas such as the yard.
- **Floors**: The number of levels the house has.
- **Waterfront**: Indicates whether the house is located on the waterfront (0 for no, 1 for yes).
- **View**: A rating of the quality of the view, ranging from 0 (no view) to 4 (excellent view).
- **Condition**: A rating of the house's overall condition on a scale of 1 (poor) to 5 (excellent).
- **Grade**: A rating of the construction and design quality, ranging from 1 (low) to 13 (high).
- **Year Built and Renovated**: The year the house was originally built and, if applicable, the year it was renovated.
- **Location**: Geospatial details such as the house's latitude, longitude, and zip code.



```
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   id             21613 non-null  int64  
 1   date           21613 non-null  object 
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64  
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64  
 6   sqft_lot       21613 non-null  int64  
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64  
 9   view           21613 non-null  int64  
 10  condition      21613 non-null  int64  
 11  grade          21613 non-null  int64  
 12  sqft_above     21611 non-null  float64
 13  sqft_basement  21613 non-null  int64  
 14  yr_built       21613 non-null  int64  
 15  yr_renovated   21613 non-null  int64  
 16  zipcode        21613 non-null  int64  
 17  lat            21613 non-null  float64
 18  long           21613 non-null  float64
 19  sqft_living15  21613 non-null  int64  
 20  sqft_lot15     21613 non-null  int64  
dtypes: float64(6), int64(14), object(1)
```

Fig.1.Dataset Information

This included handling missing values, removal of redundant features, and scaling of numerical attributes to handle skewed distributions. Now, this cleaned dataset gives the proper foundation for analysis and predictive modeling.

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Prices

The distribution of house prices is shown on Fig. 2. The data shows the right skewed pattern with most properties having prices below $1,000,000. A few properties with extremely high prices act as outliers, which contributes to the skewness. This pattern needs scaling techniques such as Robust Scaler to mitigate the impact of outliers during model training.

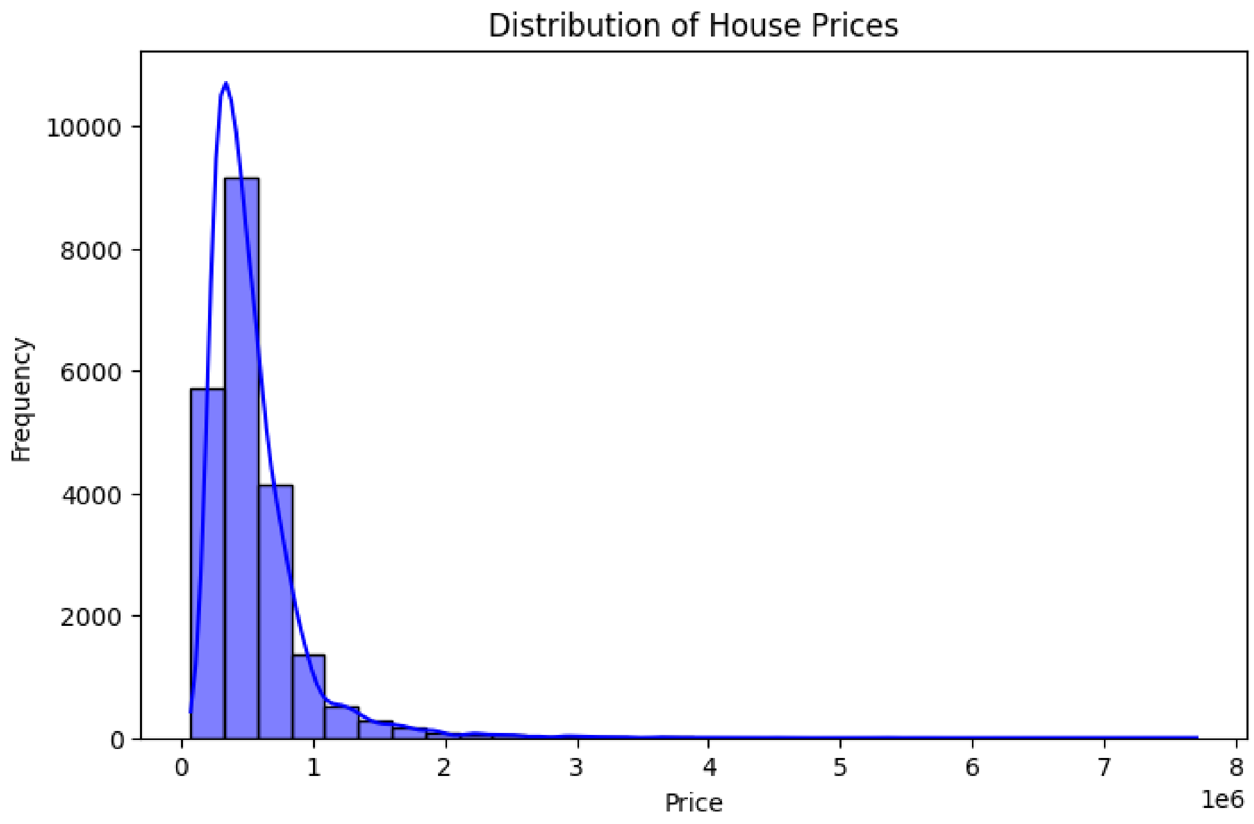


Fig. 2. Distribution of House Prices

4.2 Distribution of Bedrooms

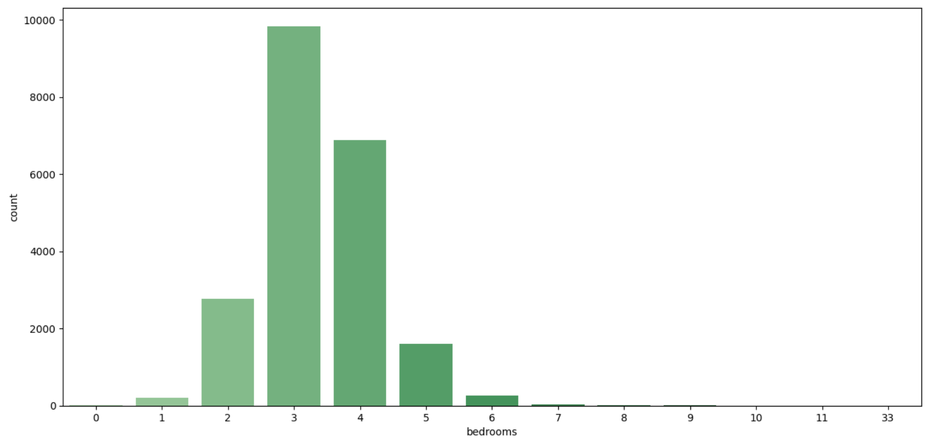


Fig. 3. Distribution of Bedrooms.

Most houses have 3 bedrooms, followed by 4 bedrooms and 2 bedrooms, while the fewest would have more than 6 bedrooms. Considering a few outliers, those up to 33 bedrooms, the trend shows that most of them are really for small to medium-sized families. (Fig.3)

4.3 Distribution of the number of floors in houses

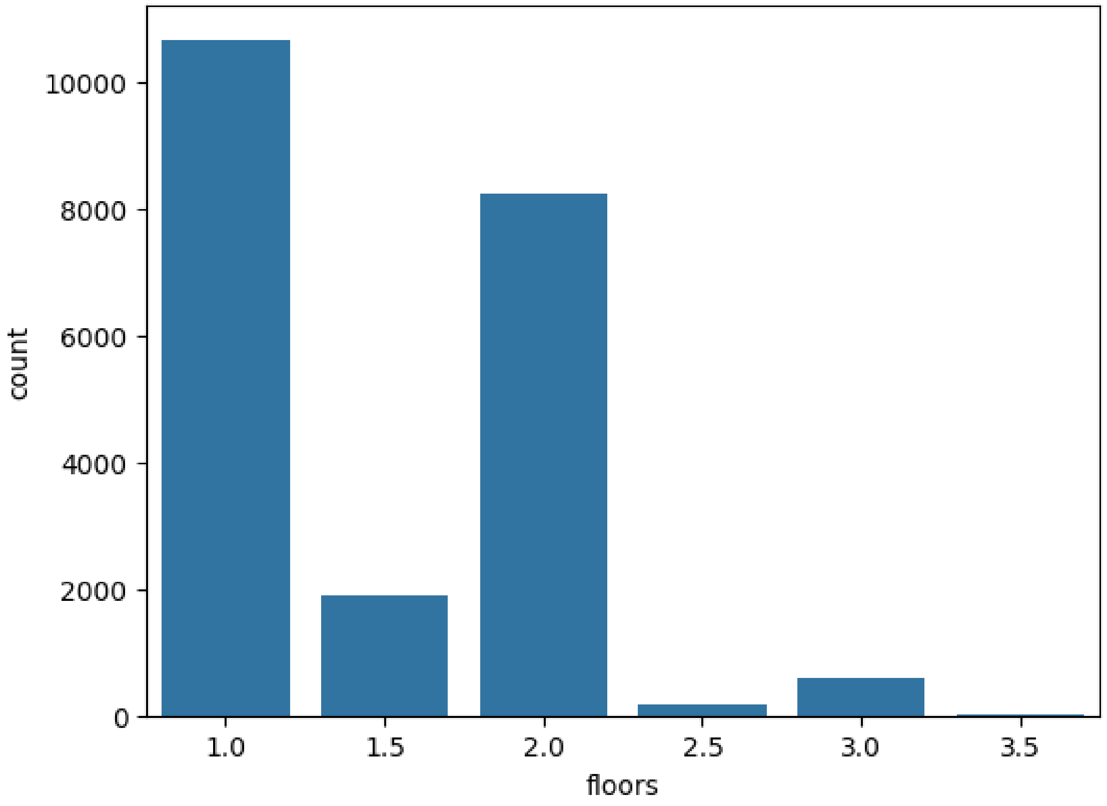


Fig. 4. Distribution of Floors.

Fig. 4 shows the distribution of the number of floors in houses. Most of the houses have either 1 or 2 floors; single-story homes are most common. There is little evidence of houses having more than two floors, representing the regional preference for compact and practical designs.

4.4 Pairplot relationships between features

A positive correlation was observed between price and sqft_living which indicates that larger living spaces generally result in higher house prices. Similarly, bathrooms and prices also show a positive relationship. (Fig.5)

The distributions (Fig.5) along the diagonal highlight that price and sqft_living are right-skewed, while bedrooms and bathrooms show more clustered distributions.

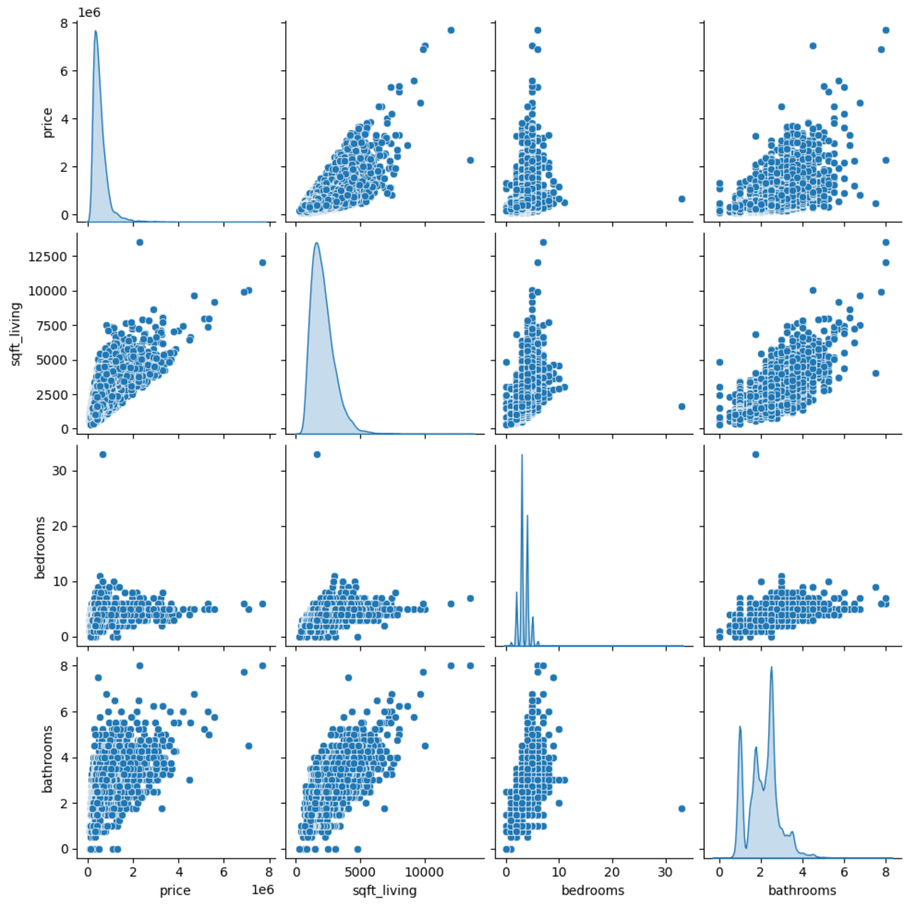
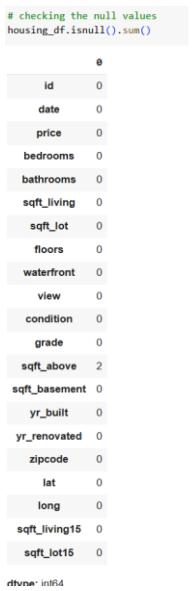


Fig. 5. Pairwise relationships between selected features.

5. Methodology

5.1 Data Preprocessing

The raw dataset was cleaned and prepared for analysis to ensure the reliability and accuracy of predictions:



```
# checking the null values
housing_df.isnull().sum()
```

```
               0
id             0
date           0
price          0
bedrooms       0
bathrooms      0
sqft_living    0
sqft_lot       0
floors         0
waterfront     0
view           0
condition      0
grade          0
sqft_above     2
sqft_basement  0
yr_built       0
yr_renovated   0
zipcode        0
lat            0
long           0
sqft_living15  0
sqft_lot15     0

dtype: int64
```

Fig. 6. Null value analysis of the dataset.

Most of the features are complete, and there are no missing values in them. Only the sqft_above feature contains 2 missing entries(Fig.6), which were addressed during the preprocessing stage by removing the affected rows. Ensuring a clean dataset was critical for improving model reliability and accuracy.
Handling Missing Values: Rows with missing entries were removed to maintain data consistency.

Feature Transformation:

- The date column was converted to the year of sales for better temporal analysis.
- Irrelevant features such as ID were dropped.

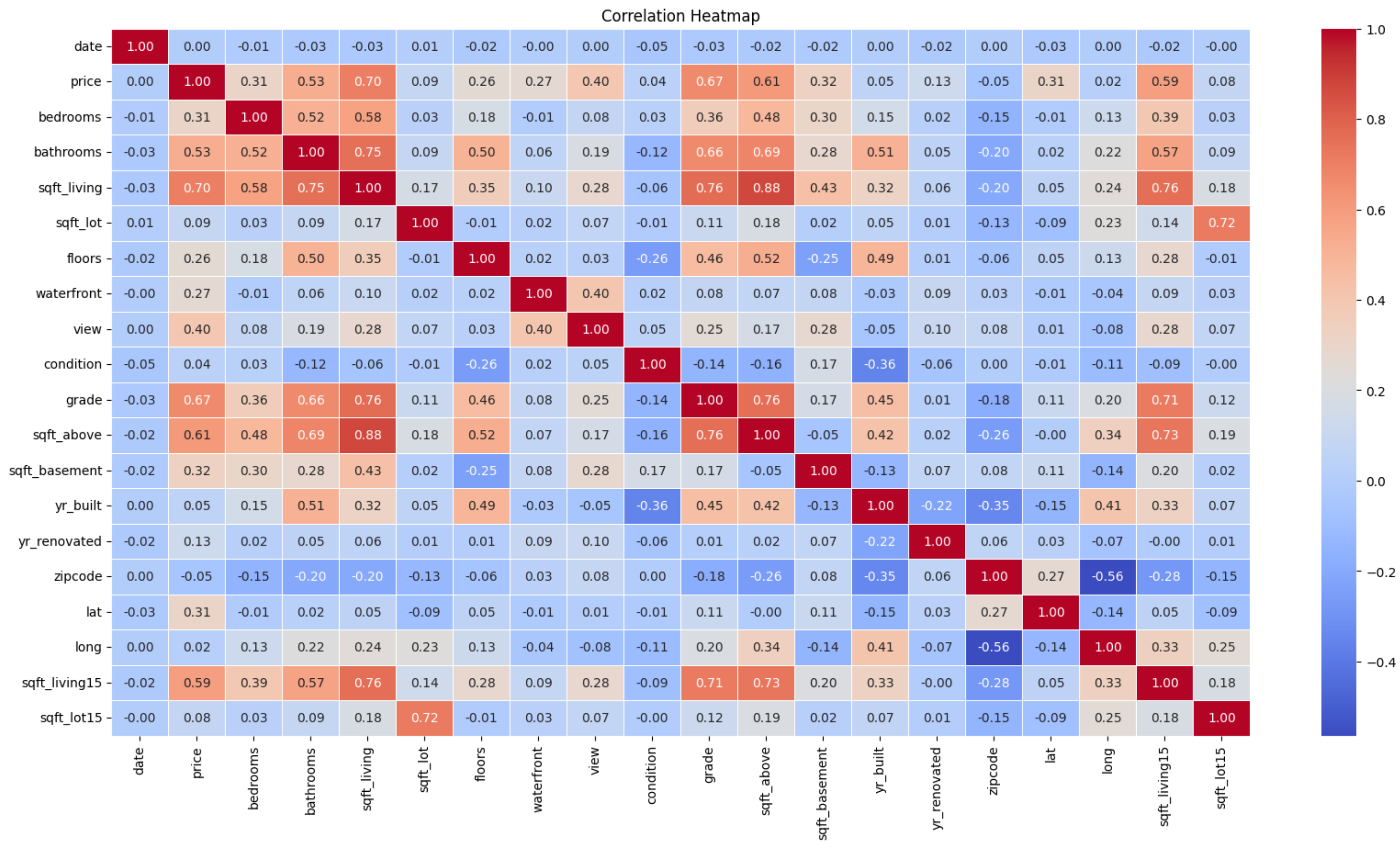


Fig.7.Heatmap

- Highly correlated features, like sqft_above (correlated with sqft_living), were removed to address multicollinearity.(Fig.7)

Feature Scaling: RobustScaler was applied to numerical features to handle skewness and outliers, making the data suitable for machine learning algorithms[2]

Robust Scaler [2] is a data preprocessing technique applied to numerical scale features. It reduces the impact of outliers because, instead of using the mean and standard deviation for scaling. It uses the median and interquartile range (IQR). Hence, this is effective for skewed data or datasets containing extreme values.

5.2 Feature Importance Analysis

To identify the most influential factors affecting house prices, feature importance[3] scores were derived using ensemble models such as Random Fores. Key predictors included sqft_living and grade.

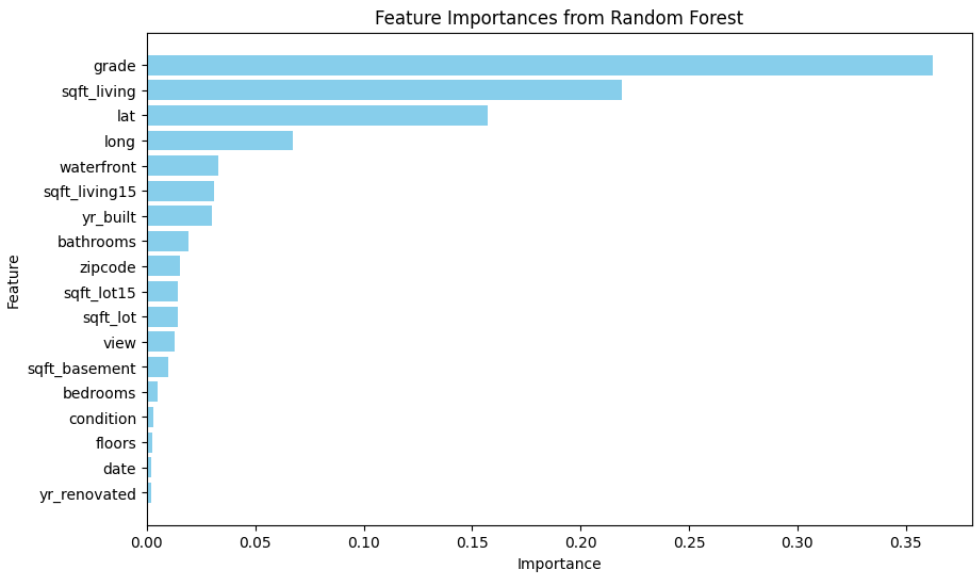


Fig. 8. Feature importances [3] from the Random Forest model.

Grade and sqft_living are the most important predictors of house prices which followed by geospatial features lat and long. The other important features are waterfront, sqft_living15, and yr_built. All these features collectively explain most of the variance in house prices. Therefore, confirm their relevance in predictive modeling. (Fig.8)

5.3 Model Development

A variety of models were developed and evaluated to identify the best performing approach for house price prediction:
Traditional Regression Models: Linear Regression, Decision Tree, and K-Nearest Neighbors.

Linear Regression: Simple, interpretable, assumes a linear relationship between the target variable and predictors. Linear regression may be used as a baseline against which to evaluate more complicated methods.

Decision Trees: Non-linear model [8] that splits data into subsets using feature thresholds. Intuitive and works well for capturing variable interactions but tends to overfit if not regularized.

K-Nearest Neighbors (KNN): A distance-based model that predicts an outcome by averaging the target values of the k-nearest data points. It is simple to implement but sensitive to data scaling and irrelevant features.[7]

Ensemble Methods: Random Forest, Gradient Boosting, and XGBoost [6] which are leveraging their capability to model complex relationships and improve accuracy.

Random Forest [4] is an ensemble method that combines numerous decision trees using random subsets of data and features to train each tree. It thus reduces overfitting while providing robust performance by averaging out the predictions across trees.

Gradient Boosting: This is a sequential ensemble technique of building models iteratively, correcting errors made by previously built models. It is very effective in handling complicated data but may be expensive in computation.[5]

XGBoost: This is an optimized version of Gradient Boosting with regularization, parallel processing, and efficient tree building. It is highly accurate and widely used for structured data tasks.[6]

Artificial Neural Network (ANN):
ANNs are inspired by the human brain and consist of interconnected layers of neurons. They excel in capturing complex, nonlinear relationships among data [9].

Architecture: This ANN model consists of three hidden layers with ReLU activation functions and dropout regularization to prevent overfitting.

Optimizer: The model uses the Adam optimizer while the loss function is Mean Squared Error(MSE).

ANNs can adapt to a wide range of data which makes them suitable for capturing intricate patterns.
Challenges: They require a lot of computational resources and tuning of hyperparameters for optimal performance.
The custom architecture was implemented with:

- Input layer: 19 features.
- Three hidden layers (128, 64, and 32 neurons) with ReLU activations.
- Dropout regularization to prevent overfitting.
- Output layer for regression with linear activation.

5.4 Ensemble Modeling (Stacking Regressor)

An ensemble model was created by taking weighted average predictions from Random Forests, Gradient Boosting, and XGBoost [6]. This is done to improve the accuracy of prediction by considering the strengths of multiple models.

5.5 Model Evaluation

The models were evaluated on validation and test datasets using the following metrics:

- **R-squared (R^2):** It measures how well the model explains the variation in the target variable. The higher the R^2, the better the model fits the data.
- **Adjusted R-squared:** This is like R^2, but it also adjusts for the number of predictors to prevent adding irrelevant features from artificially inflating the value.
- **Mean Squared Error:** The average of squared differences between forecasted and actual values. It is a global measure of error.
- **Root Mean Squared Error (RMSE):** The square root of MSE, in the same units as the target variable. Hence, it is more interpretable.
- **Mean Absolute Error (MAE):** It is the average of absolute differences between predicted and actual values which provide an intuitive measure of the inaccuracy of predictions.

6. Results

6.1 Model Performance

The models were evaluated on their predictive ability using metrics such as R-squared (R^2), Adjusted R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Model	R^2	Adjusted R^2	MSE	RMSE	MAE
Linear Regression	0.7111	0.7098	4.32×10^{10}	207,769	127,177
K-Nearest Neighbors	0.7122	0.711	4.29×10^{10}	207,350	103,042
Random Forest	0.8713	0.8707	1.92×10^{10}	138,671	71,901
Gradient Boosting	0.8696	0.869	1.94×10^{10}	139,597	80,508
Decision Tree	0.7526	0.7516	3.69×10^{10}	192,246	104,350
XGBoost	0.8714	0.871	1.91×10^{10}	138,427	68,961
ANN	0.7623	0.7613	3.55×10^{10}	188,459	107,870

6.2 Model Performance Comparison:

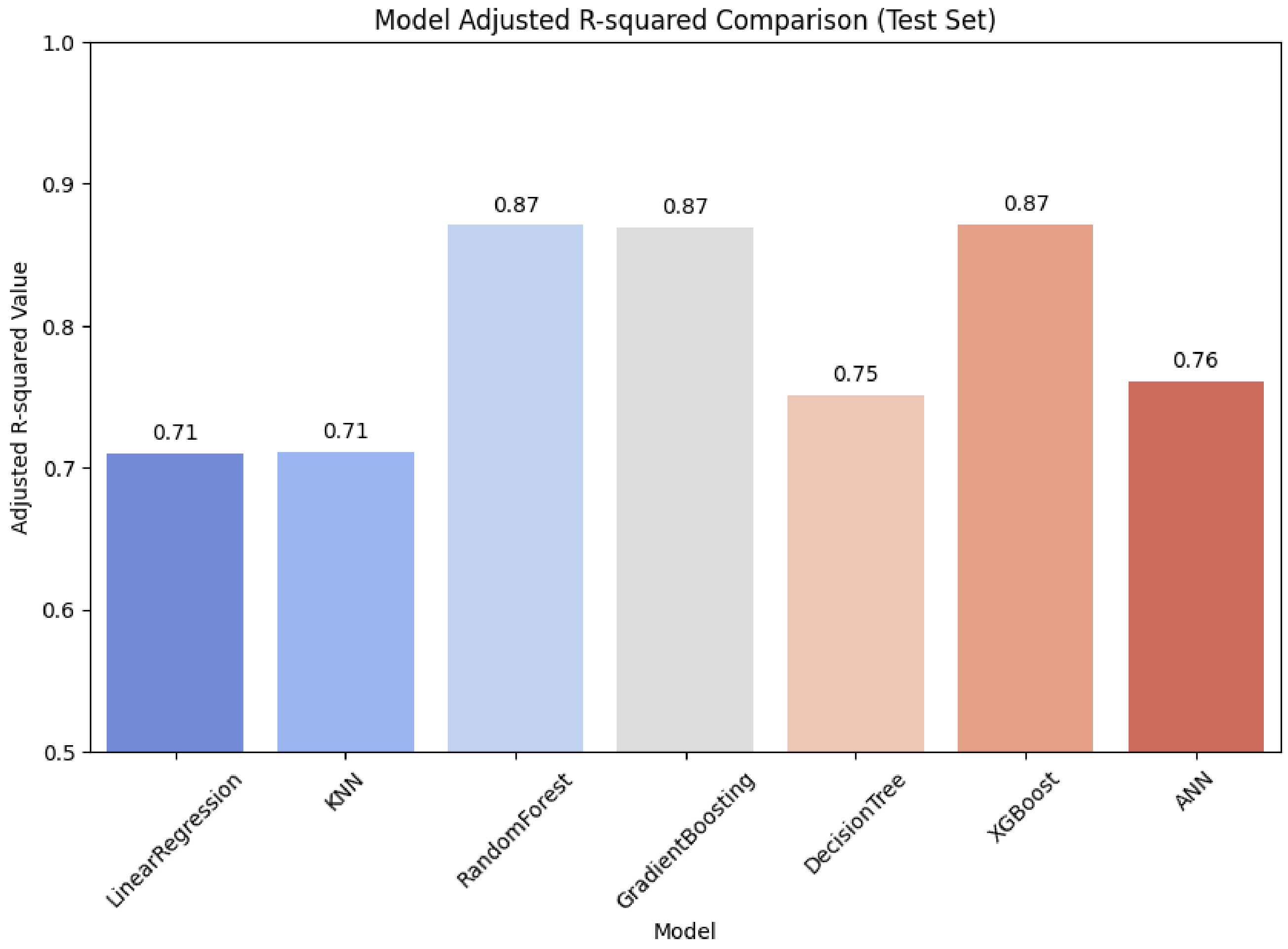


Fig. 8. Adjusted R-squared comparison of models on the test set.

The best results in terms of the Adjusted R-squared were obtained with Random Forest, Gradient Boosting, and XGBoost [6] with a value of 0.87 which reflects their strong predictive capabilities adjusted for model complexity. Linear Regression and KNN performed comparatively with an Adjusted R-squared of 0.71. The Decision Tree model reached a moderate value of 0.75, while the ANN [9] had an Adjusted R-squared of 0.76. (Fig.8)

7. Discussion

Best-Performing Model: The XGBoost model is the best in performance. it has the highest R^2 (0.8714) and the lowest MAE (68,961). This depicts how good the model is in house price prediction.

Ensemble models: such as Random Forest [4] and Gradient Boosting [5], performed very well and showed very competitive R^2 values with lower errors compared to other models.

Artificial Neural Networks (ANNs) performed well but were less accurate compared to tree-based ensemble models. Improving their design could lead to better results.[9]
Model Optimization:

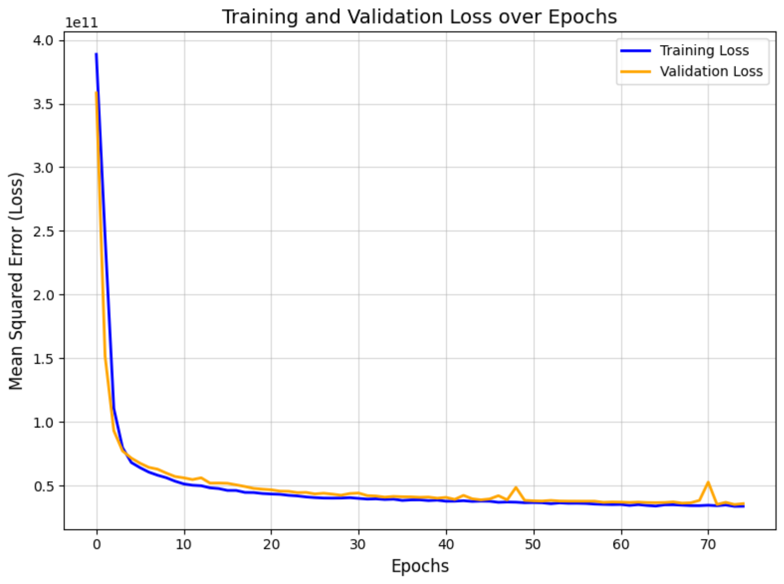


Fig.9.Training and Validation Loss over Epochs for ANN Model

It significantly decreases within the first few epochs and then stabilizes at approximately epoch 30. Convergence of training and validation loss is indicative of good generalization and minimal overfitting.

Key Predictors: Feature importance analysis identified sqft_living, grade, and bathrooms as the most important factors that influence house prices which correspond to domain knowledge.
The result proves that ensemble models, especially XGBoost are good at predicting house prices because they can handle the complex relations in the data.

Ensemble Model (Stacking Regressor) Performance:

The Stacking Regressor [10] which combines predictions from XGBoost, Random Forest, and Gradient Boosting achieved the best performance with the highest R^2 (0.8762). This demonstrates the effectiveness of leveraging multiple models for enhanced accuracy.

8. Conclusion

This project successfully developed and evaluated multiple predictive models for house price estimation using the King County Housing Dataset. One was able to conclude that XGBoost [6] is one of the most skilled. Additionally, models such as Random Forest [4] and Gradient Boosting [5] turned out to be reliable alternatives for ensembles.

Artificial Neural Networks (ANNs) have demonstrated considerable promise. However, additional advancements are necessary to optimize. Findings derived from the analysis of feature importance offer significant insights into the determinants that affect property values, thereby facilitating their application in real world scenarios. Future work could be investigated: Improving ANN architecture [9] with such techniques as batch normalization and dropout optimization. Integrate more external datasets to enhance model generalization. Use advanced ensemble techniques to improve the accuracy of predictions.

9. References

[1] King County, USA. (n.d.). King County Housing Dataset. https://www.kaggle.com/datasets/shivachandel/kc-house-data/data

[2] Scikit-learn Developers. (2023). **RobustScaler**. Link: https://scikit-learn.org.

[3] Scikit-learning Developers. (2023). **Feature Importance in Ensemble Models**. Link: https://scikit-learn.org.

[4] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[5] Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378.

[6] Chen, T., & Guestrin, C. (2016). **XGBoost: A Scalable Tree Boosting System**. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[7] Zhang, H., & Wang, D. (2006). **K-Nearest Neighbors and Applications**. *Handbook of Research on Machine Learning Applications and Trends*.

[8] Quinlan, J. R. (1986). **Induction of Decision Trees**. *Machine Learning*, 1(1), 81-106.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). **Deep Learning**. MIT Press, Comprehensive coverage of artificial neural networks and their architectures.

[10] Scikit-learn Developers. (2023). **Stacking Regressor Documentation**. Link: https://scikit-learn.org