1. Introduction

The real estate market plays a critical role in every economy, influencing investment decisions, economic policies, and individual financial planning. Accurate prediction of house prices is crucial for various stakeholders, including buyers, sellers, real estate agents, and policymakers. The objective of this project is to develop a predictive model that can forecast house prices using a dataset of American housing data. By leveraging advanced regression techniques, we aim to build a model that provides high accuracy and reliability in predicting home sales prices.

2. Problem Statement

Predicting house prices involves understanding and analysing numerous factors influencing property values, such as the number of bedrooms and bathrooms, living space, location-specific demographics, and economic indicators. The complexity of the problem requires sophisticated modelling techniques to capture the nuances in the data and provide reliable price predictions.

This project aims to address the following questions:

What are the significant factors influencing house prices?

How accurately can we predict house prices using advanced regression techniques?

What trends and insights can be derived from the data to aid decision-making in the real estate market?

3. Dataset Description

The dataset used for this project is sourced from Kaggle and contains detailed information on house prices in the American real estate market. The dataset includes the following features:

Zip Code: The postal code of the property.

Price: The sale price of the house.

Beds: Number of bedrooms.

Baths: Number of bathrooms.

Living Space: Square footage of the living space.

Address: Physical address of the property.

City: The city where the property is located.

State: State where the property is located.

Zip Code Population: Population in the zip code area.

Zip Code Density: Population density in the zip code area.

County: The county where the property is located.

Median Household Income: Median household income in the area.

Latitude: Geographic latitude.

Longitude: Geographic longitude.

Initial exploration of the dataset revealed the need for data cleaning and preprocessing, including handling missing values, removing duplicates, and feature engineering to create new variables that might enhance the model's predictive power.

4. Data Preprocessing

Data preprocessing is crucial to ensure the dataset is clean, consistent, and ready for modelling. The following steps were undertaken:

Removal of Duplicates: Duplicate entries were removed to maintain data integrity, reducing the dataset to 39,019 rows and 14 columns.

Handling Missing Values: The Median Household Income column had two missing values, which were filled with the median value of the column to preserve the statistical distribution.

Feature Engineering: A new feature, Price per SqFt, was created to provide better insight into property values. This feature is calculated as the ratio of the house price to the living space area.

The cleaned dataset was then split into training and testing sets, with 80% of the data used for training the models and 20% reserved for testing.

5. Algorithm Selection

Various regression algorithms were considered for this project:

Linear Regression: A basic approach modelling the relationship between features and the target variable as a linear combination (Neter et al., 1996).

Ridge Regression: An extension of linear regression that includes L2 regularisation to prevent overfitting by penalising large coefficients (Tibshirani, 1996).

Lasso Regression: Similar to ridge regression but uses L1 regularisation to prevent overfitting and perform feature selection by shrinking some coefficients to zero (Tibshirani, 1996).

Gradient Boosting Regression: An ensemble technique that builds trees sequentially, where each tree corrects errors from the previous ones (Friedman, 2001).

These algorithms provide a mix of linear and ensemble approaches, allowing us to compare their performance and select the best model for predicting house prices.

6. Model Training and Evaluation

The models were trained on the training set and evaluated on the testing set using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as performance metrics.

Linear Regression

The Linear Regression model served as a baseline. It showed high errors, indicating it might not capture the complexity of the data.

Training Set:

MAE: $277,517.25

RMSE: $686,251.07

Testing Set:

MAE: $287,725.84

RMSE: $881,271.11

Ridge Regression

Ridge Regression performed similarly to Linear Regression, indicating that overfitting was not a major issue in the linear model.

Training Set:

MAE: $277,516.58

RMSE: $686,251.07

Testing Set:

MAE: $287,725.08

RMSE: $881,271.08

Lasso Regression

Lasso Regression also showed similar performance, suggesting that feature selection did not significantly impact the results.

Training Set:

MAE: $277,516.96

RMSE: $686,251.07

Testing Set:

MAE: $287,725.52

RMSE: $881,271.18

Gradient Boosting Regression

Gradient Boosting Regression performed well, capturing complex relationships in the data.

Training Set:

MAE: $25,677.34

RMSE: $52,395.72

Testing Set:

MAE: $30,399.71

RMSE: $125,123.16

7. Results

The Gradient Boosting Regression model emerged as the best performer, showing significantly lower errors compared to other models. Its ability to handle non-linear relationships and its ensemble nature made it the most suitable choice for this task. The detailed performance metrics are:

Training Set:

MAE: $25,677.34

RMSE: $52,395.72

Testing Set:

MAE: $30,399.71

RMSE: $125,123.16

The model's predictions provide valuable insights into the factors influencing house prices and can aid stakeholders in making informed decisions.

The tables below illustrate the MAE and RMSE

### **Mean Absolute Error (MAE)**

| **Model** | **Train MAE ($)** | **Test MAE ($)** |
| --- | --- | --- |
| Linear Regression | 277,517.25 | 287,725.84 |
| Ridge Regression | 277,516.58 | 287,725.08 |
| Lasso Regression | 277,516.96 | 287,725.52 |
| Gradient Boosting Regression | 25,677.34 | 30,399.71 |

### 

### 

### 

### **Root Mean Squared Error (RMSE)**

| **Model** | **Train RMSE ($)** | **Test RMSE ($)** |
| --- | --- | --- |
| Linear Regression | 686,251.07 | 881,271.11 |
| Ridge Regression | 686,251.07 | 881,271.08 |
| Lasso Regression | 686,251.07 | 881,271.18 |
| Gradient Boosting Regression | 52,395.72 | 125,123.16 |

8. Conclusion

The project successfully developed a predictive model for house prices using advanced regression techniques. The Gradient Boosting model outperformed other algorithms, providing accurate and reliable predictions. Key observations from the project include:

Significant Factors: Features such as living space, number of bedrooms and bathrooms, and median household income significantly impact house prices.

Model Performance: Ensemble models like Gradient Boosting offer superior performance due to their ability to handle complex relationships and reduce overfitting.

Future Trends: The model can be used to predict future house prices, helping stakeholders to make data-driven decisions.

Future work includes further feature engineering, hyperparameter tuning of the Gradient Boosting model, and incorporating more recent and comprehensive datasets to enhance the model's accuracy and relevance.

9. References

Kaggle. "American Housing Data". Available at: American Housing Data.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models. McGraw-Hill.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58(1), 267-288. Available at: Lasso Regression Paper.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Available at: Scikit-learn Documentation.