House Price Prediction
Progress Report

Nishant Yadav
2022329
nishant22329@iiitd.ac.in

Kumar Mrinal
2022258
mrinal22258@iiitd.ac.in

Aditya Kumar
2021124
aditya21124@iiitd.ac.in

Abstract

The real estate market significantly impacts the economy, influencing financial decisions and broader trends. Accurate house price predictions benefit stakeholders like homeowners, buyers, investors, and financial institutions. Traditional methods often fall short in capturing the complex interplay of factors such as location, property traits, market trends, and economic conditions. Machine learning offers a powerful alternative by analyzing large datasets to uncover hidden patterns. This project utilizes machine learning to enhance house price prediction accuracy, aiding decision-making in the housing market.

1. Introduction

This project aims to develop a machine learning model to predict house prices using features like size, number of rooms, construction quality, and year built. The dataset [1] includes both numerical and categorical data. Initial modeling will use linear regression as a baseline, followed by advanced methods like Random Forest Regressor, Support Vector Machines (SVM) and boosting techniques to enhance accuracy. By analyzing features, training models, and evaluating performance, the project aims to uncover insights into price determinants and provide a tool for informed decision-making in the real estate market.

Through feature analysis, model training, and evaluation, this project seeks to provide valuable insights into how various features influence house prices, and ultimately develop a model that can assist stakeholders in making informed decisions in the real estate market.

2. Literature Survey

Predicting house prices has been extensively studied, with traditional methods like linear regression commonly used to analyze relationships between factors such as size, bedrooms, and location. However, these approaches struggle with capturing non-linear relationships.

Advancements in machine learning introduced models like Decision Trees, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), offering improved performance by handling complex patterns. Thuraiya et al. (2020) [2] demonstrated Random Forest's superior accuracy over regression due to reduced variance. Similarly, Li et al. (2009) [3] showed SVM outperformed linear regression for non-linear relationships, though careful hyperparameter tuning was necessary.

Varma et al. (2018) [4] explored neural networks, highlighting their high accuracy with large datasets but noting their computational expense and preprocessing requirements. Lu et al. (2017) [5] emphasized feature selection, showing that factors like neighborhood quality and amenities significantly enhance predictions. Ensemble methods like Gradient Boosting and XGBoost (Rana et al. (2020) [6])further improved accuracy by capturing variable interactions in noisy datasets. This project builds on these studies, evaluating linear regression, SVMs, and other models for house price prediction using real-world data.

3. Dataset

3.1. Dataset Overview

The dataset used in this project is based on house sales in Ames, Iowa. It contains detailed information about various aspects of each property, including structural features, lot characteristics, neighborhood location, and sale conditions.

Here's a brief description of some important features:

- GrLivArea: Above grade (ground) living area square feet
- TotalBsmtSF: Total square feet of basement area
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- GarageArea: Size of garage in square fee

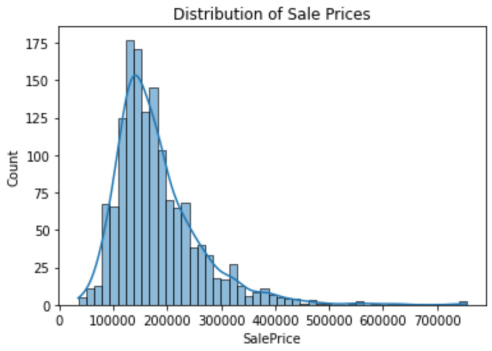

Figure 1. Distribution of SalePrice

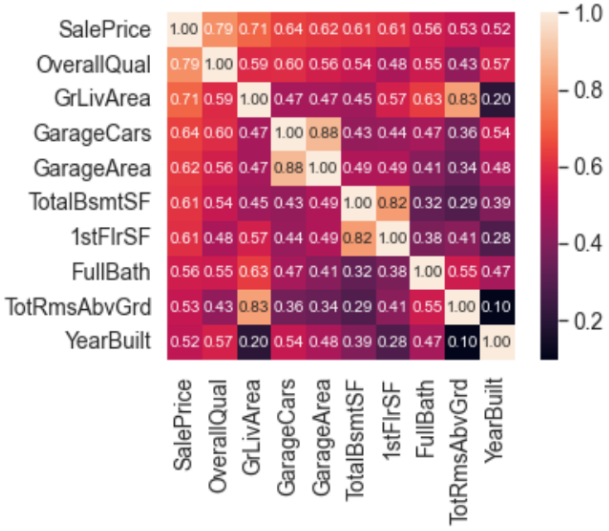

Figure 2. Correlation heatmap of top ten numerical features

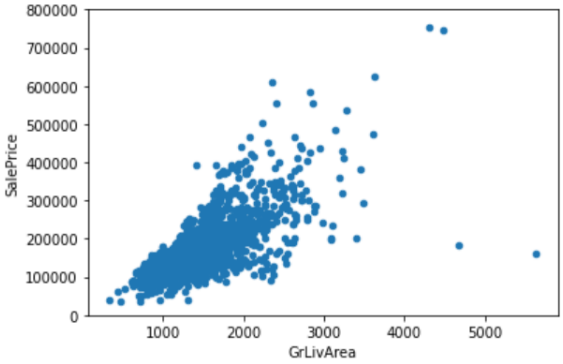

Figure 3. Above Grade Living Area vs Sale Price

There are over 80 features in the dataset that contain information about various other aspects of houses. The detailed description of the dataset is available here: `https://github.com/y-nishant/House-Price-Prediction/blob/main/data_description.txt`.

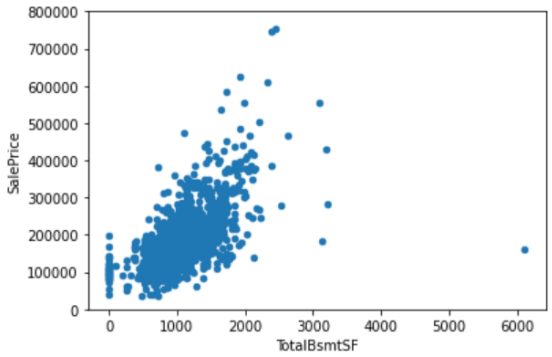

Figure 4. Basement Area vs Sale Price

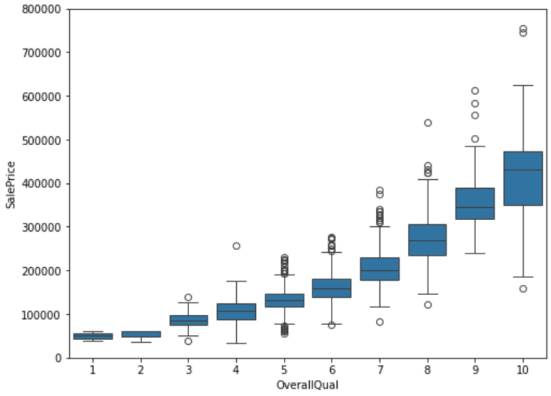

Figure 5. Overall Quality vs Sale Price

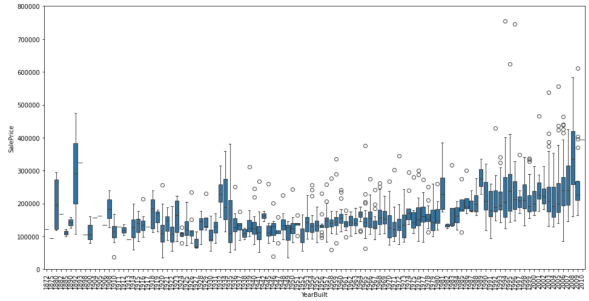

Figure 6. Year Built vs Sale Price

3.2. Data Visualization

From Figure 1, we can observe that SalePrice deviates from the normal distribution, it shows positive skewness and peakedness.

In the correlation heatmap (Figure 2), we can observe that Overall Quality, Above Grade Living Area, Garage Area, Total Basement Area and First Floor Area are highly correlated to Sale Price.

Above Grade Living Area (GrLivArea), Basement Area (TotalBsmtSF) and Overall Quality (OverallQual) show a linear relationship with respect to Sale Price as observed

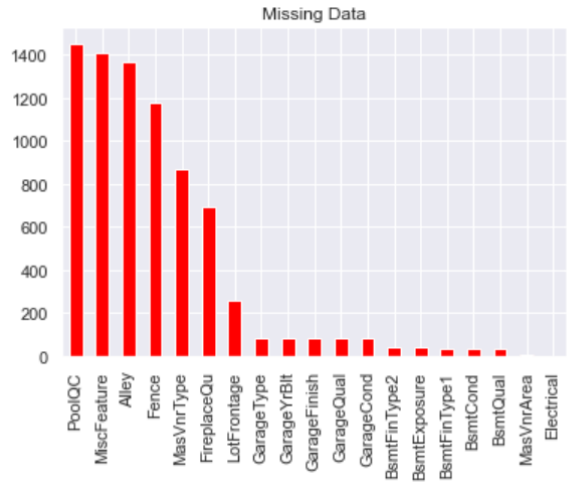

Figure 7. Missing data for each feature

from Figure 3, Figure 4 and Figure 5. We can observe the same linear relationship for Year Built with respect to Sale Price (Figure 6), however we do not know if 'SalePrice' is constant price or not. Constant prices remove the effect of inflation. If 'SalePrice' is not in constant prices, Sale Prices are not fairly comparable over the years.

3.3. Data Preprocessing

Handling Missing Values: Features that had a lot of missing values were dropped from the dataset - 'PoolQC', 'MiscFeature', 'Alley' and 'Fence'
For numerical features, missing values were imputed using the median value of the respective feature.
Categorical features were filled with a special category "NA" to indicate missing information.

Encoding Categorical Features: Categorical features were converted into numerical form using one-hot encoding. This allows the machine learning model to interpret them effectively.

Scaling Features: Continuous numerical features were scaled using MinMax Scaling technique to ensure that features with larger ranges did not dominate during model training.

Feature Engineering: Polynomial features of degree 2 were generated using PolynomialFeatures to capture non-linear relationships in the data.

Model Tuning: Hyperparameter optimization was performed using GridSearchCV. A grid search over parameters such as n-estimators (100, 200) and max-depth (None, 10, 20) was conducted with 5-fold cross-validation. The best model was selected based on the lowest negative mean squared error.

4. Methodology

Several machine learning models were chosen for training, covering both linear and non-linear approaches to regression. The models were trained using the training data, and their performances were evaluated based on the error metrics discussed in the evaluation section.

4.1. Linear Regression

Linear Regression is a fundamental regression algorithm that assumes a linear relationship between the input features and the target variable. It was used as a baseline model to assess the performance of more complex models.

$$\min_{\beta} \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 \tag{1}$$

4.2. Support Vector Regression (SVR)

SVR is a more advanced technique that tries to find a hyperplane that fits the data while maintaining a margin of tolerance (epsilon). It can capture non-linear relationships between features and the target variable.

The main task of SVR is to find the best hyperplane (point in case of one dimensional, line in case of two dimensional, a plane in case of three dimensional and hyperplane in case of n-dimensional) which linearly separates the data points into two-component by maximizing the margin.

4.3. Elastic Net Regression

ElasticNet is a regularized linear model that combines both L1 (Lasso) and L2 (Ridge) penalties. This helps in controlling overfitting by constraining the model coefficients, especially in datasets with multicollinearity or many features.

$$\min_{\beta} \quad \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \tag{2}$$

4.4. Random Forest Regressor

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the average prediction of the individual trees to improve accuracy and reduce overfitting. This method is particularly useful for capturing complex interactions between features.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{3}$$

4.5. XGBoost

XGBoost (Extreme Gradient Boosting) is a popular and efficient implementation of gradient boosting. It builds an ensemble of decision trees, where each tree corrects the errors of its predecessor. XGBoost uses a technique called "boosting" to sequentially add trees that focus on the mistakes made by previous ones. It incorporates regularization (L1 and L2) to prevent overfitting and uses advanced features like column subsampling, tree pruning, and handling missing values automatically. XGBoost has been shown to provide high predictive accuracy, especially when dealing with large and noisy datasets, making it a strong choice for house price prediction.

4.6. Stochastic Gradient Descent Regressor (SGD)

Stochastic Gradient Descent is an optimization algorithm used for training linear models, particularly when working with large datasets. It updates model parameters iteratively based on the gradient of the loss function with respect to each data point, rather than using the entire dataset at once. This allows for faster convergence and better scalability compared to batch gradient descent. The SGD Regressor can incorporate regularization techniques, such as L2 (Ridge) or L1 (Lasso), to reduce overfitting by penalizing large coefficients. This makes it particularly useful for datasets with many features or when computational efficiency is a priority.

4.7. Ridge Regrerssion

Ridge Regression is a linear model that incorporates L2 regularization. It penalizes the squared magnitude of the model coefficients, thus shrinking them towards zero. This helps to address issues of multicollinearity (where features are highly correlated) and reduces overfitting by preventing the model from relying too heavily on any single feature. Ridge Regression is particularly useful when the dataset has many correlated variables or when it's important to maintain all features in the model. The regularization strength is controlled by a hyperparameter, which must be tuned to balance bias and variance effectively.

4.8. Bayesian Ridge Regrerssion

Bayesian Ridge Regression extends Ridge Regression by introducing a probabilistic approach. Instead of finding a single value for each coefficient, Bayesian Ridge estimates a distribution over possible values for the coefficients, incorporating prior knowledge and uncertainty. This approach allows the model to better account for uncertainty in the data, which is particularly valuable when working with noisy or sparse data. It also enables the model to automatically adjust the regularization strength through a process of Bayesian inference. This makes Bayesian Ridge Regression particularly useful when it's important to model uncertainty and improve the robustness of the predictions.

5. Results and Analysis

After evaluating the models, their performances were compared based on the calculated error metrics (MSE, RMSE and MAE). This comparison allowed us to determine which model provided the most accurate predictions for house prices. Additionally, model complexity, training time, and interpretability were considered when selecting the final model for deployment.

Model	MSE	RMSE	MAE
Linear Regression	1.12×10^{30}	1.06×10^{15}	1.13×10^{14}
SVR	7.18×10^{9}	84754.2	57181.6
Elastic Net	2.45×10^{9}	49514.2	30173.1
Random Forest	7.21×10^{8}	26865.4	16576.2
XGBoost	6.84×10^{8}	26164.8	17069.1
SGD	9.67×10^{8}	31082.5	20326.9
Ridge	8.17×10^{8}	28593.6	18910.5
Bayesian Ridge	8.65×10^{8}	29419.3	19380.4

Table 1. Results

We can observe that XGBoost has the lowest RMSE. Linear regression is the worst followed by SVR and Elastic Net. We can conclude that from the above models used to train the dataset, XGBoost is the best model to predict house prices.

6. Conclusion

This project on House Price Prediction has been an enriching learning experience, allowing us to apply theoretical machine learning concepts to a real-world dataset. We gained hands-on experience in data preprocessing, model training, and evaluation, deepening our understanding of various regression techniques.

7. Contribution

Nishant: Took the lead in the overall project management, conducted extensive data preprocessing (handling missing values, encoding features, hyperparameter tuning, and scaling), and implemented various machine learning models.

Mrinal: Contributed to the exploration and analysis of the dataset, performed feature engineering to improve model performance, and assisted in the evaluation and comparison of different models.

Aditya: Assisted in reviewing literature on house price prediction methodologies, supported the creation of visualizations for data analysis, and helped in the identification of key features influencing house prices.

References

[1] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.

[2] Thuraiya Mohd, Nur Syafiqah Jamil, Noraini Johari, Lizawati Abdullah, and Suraya Masrom. An overview of real estate modelling techniques for house price prediction. In *Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3 International Conference on the Future of ASEAN (ICoFA) 2019—Volume 1*, pages 321–338. Springer, 2020.

[3] Da-Ying Li, Wei Xu, Hong Zhao, and Rong-Qiu Chen. A svr based forecasting approach for real estate price prediction. In *2009 International conference on machine learning and cybernetics*, volume 2, pages 970–974. IEEE, 2009.

[4] Ayush Varma, Abhijit Sarma, Sagar Doshi, and Rohini Nair. House price prediction using machine learning and neural networks. In *2018 second international conference on inventive communication and computational technologies (ICICCT)*, pages 1936–1939. IEEE, 2018.

[5] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh. A hybrid regression technique for house prices prediction. In *2017 IEEE international conference on industrial engineering and engineering management (IEEM)*, pages 319–323. IEEE, 2017.

[6] Vivek Singh Rana, Jayanto Mondal, Annu Sharma, and Indu Kashyap. House price prediction using optimal regression techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 203–208. IEEE, 2020.