House Price Prediction Using Machine Learning

MINI PROJECT

MEHDI CHITSAZ 40132819 AND MUSTAFA ALAWADI 40217764

TO: INSTRUCTOR YASER ESMAEILI SALEHANI

DUE APRIL 11, 2025

CONCORDIA UNIVERSITY

Abstract

This project applies machine learning techniques to predict house prices using the Ames Housing Dataset. We implement data preprocessing, exploratory data analysis (EDA), and regression models to identify key housing attributes influencing prices. By comparing multiple regression algorithms, we aim to select the most accurate predictive model. The findings will help buyers, sellers, and investors make informed decisions based on data-driven insights.

Introduction

The real estate market is influenced by numerous factors, including location, property size, and structural quality. Predicting house prices is a challenging problem due to the complex relationships between these variables. Traditional appraisal methods often rely on heuristics and expert judgment, which can introduce bias and reduce accuracy. Machine learning provides a data-driven approach to price prediction by identifying patterns in historical sales data. In this project, we leverage machine learning algorithms to develop a predictive model for house prices based on the Ames Housing Dataset. By comparing multiple regression techniques, we aim to determine the most effective method for accurate price estimation.

Dataset Collection

Dataset: Ames Housing Dataset (Kaggle)

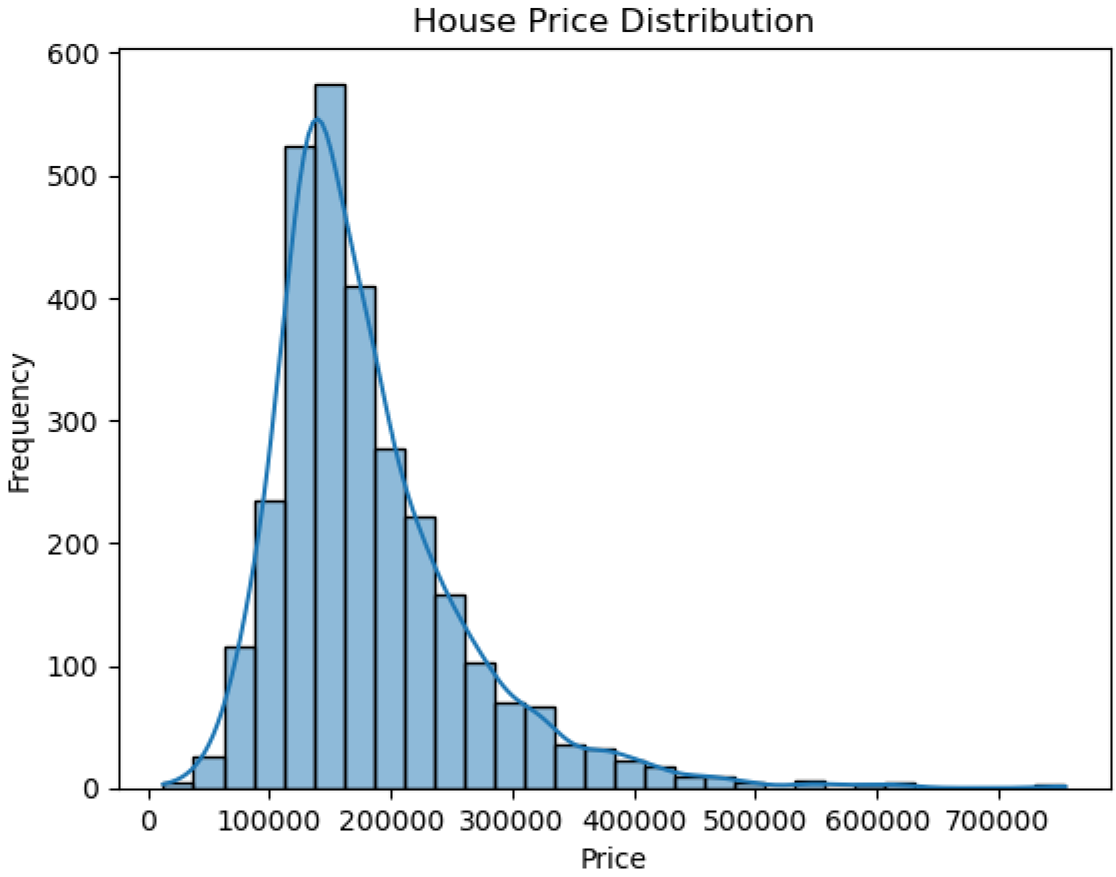
- Size: ~2,930 records, 81 features
- Target Variable: SalePrice (House price in USD)
- Key Features:
    - OverallQual: House quality rating
    - GrLivArea: Living area (sq. ft.)
    - TotalBsmtSF: Basement area (sq. ft.)
    - GarageCars: Number of garage spaces

Methodology

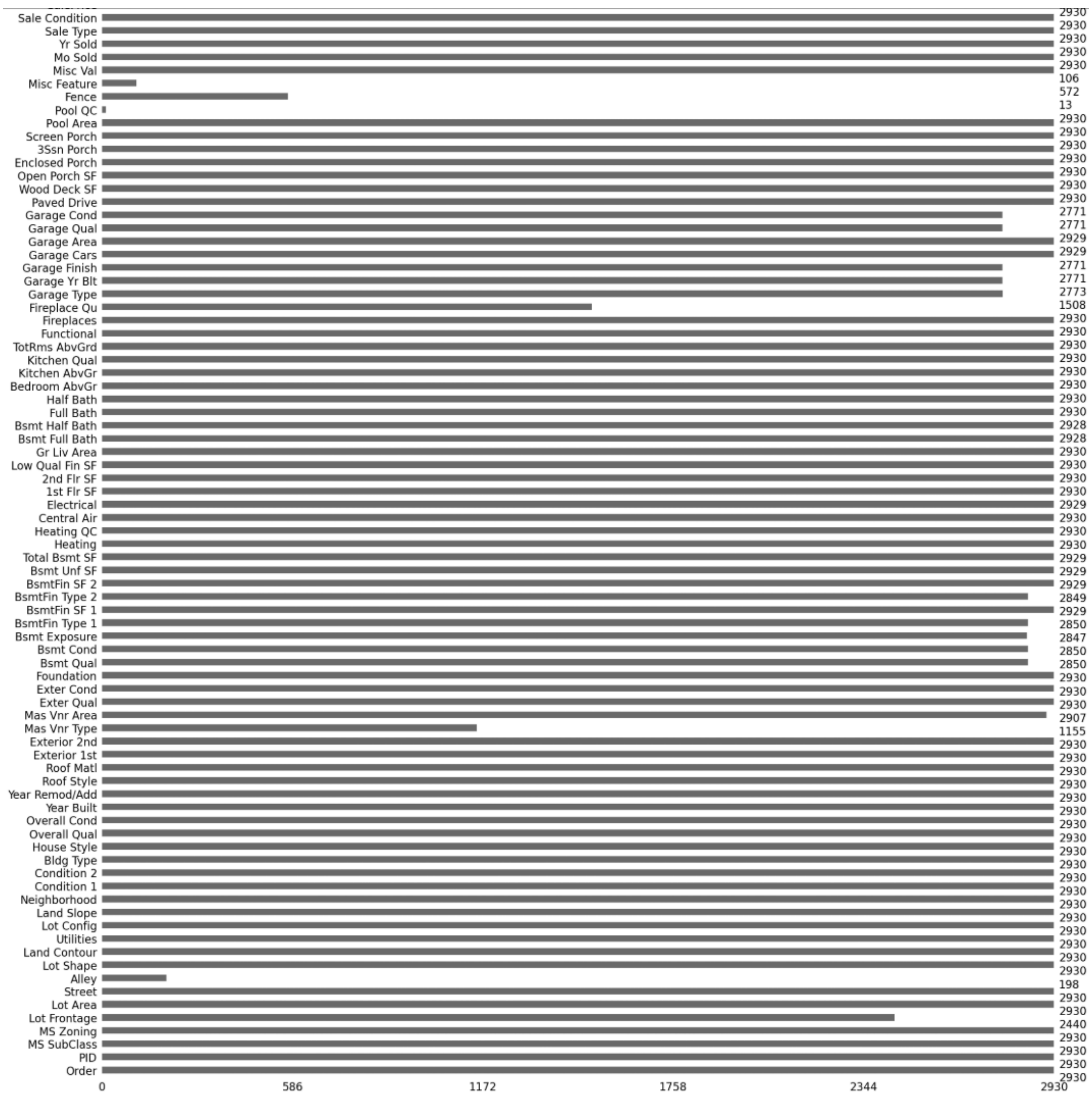
1. **Data Preprocessing & Cleaning:**

- Handling Missing Values:
  - Dropped columns with more than 50% missing values.
  - Replaced missing numerical values with the median.
  - Replaced missing categorical values with the mode.
- Encoding Categorical Variables:
  - Applied one-hot encoding to categorical features.
- Feature Selection:
  - Computed correlation between features and SalePrice.
  - Removed highly collinear variables using Variance Inflation Factor (VIF).

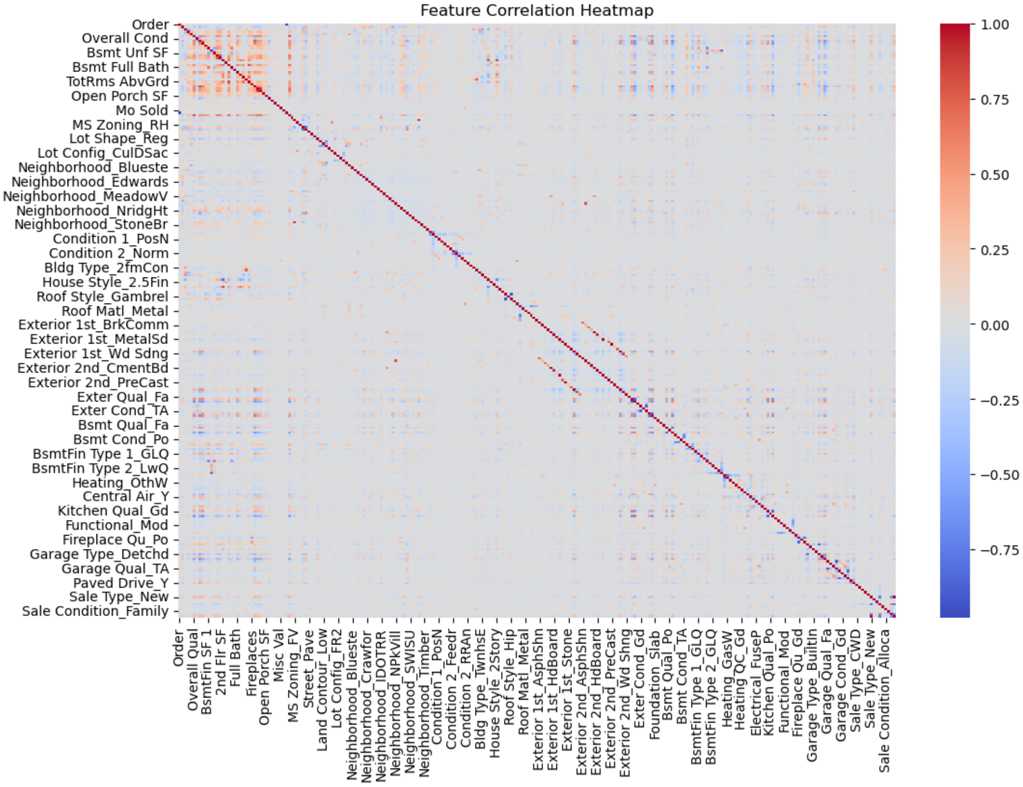
2. **Exploratory Data Analysis (EDA):**



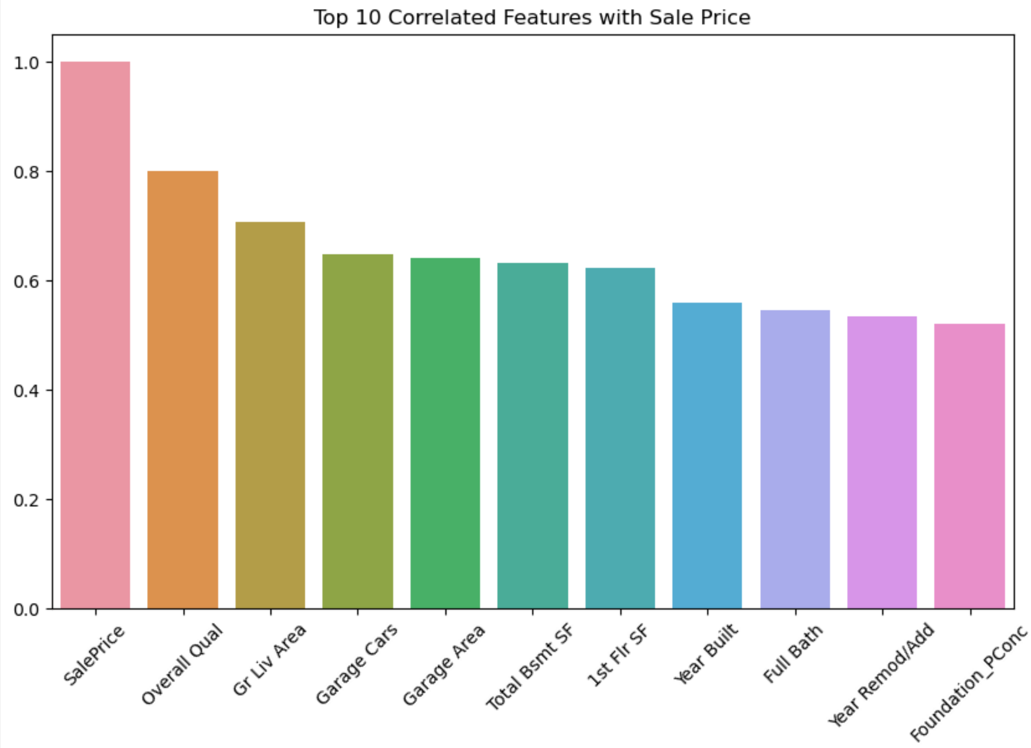
The histogram shows that house prices are right-skewed, indicating the presence of high-priced outliers.



A missing value bar chart was used to analyze and confirm the extent of missing data.



Features such as OverallQual and GrLivArea show strong positive correlation with SalePrice.



The top 10 influential factors highlight the importance of property quality, size, and location in price determination.

3. **Model Selection & Training:**

We compared three regression models:

- Linear Regression
- Random Forest Regression
- Gradient Boosting Regression

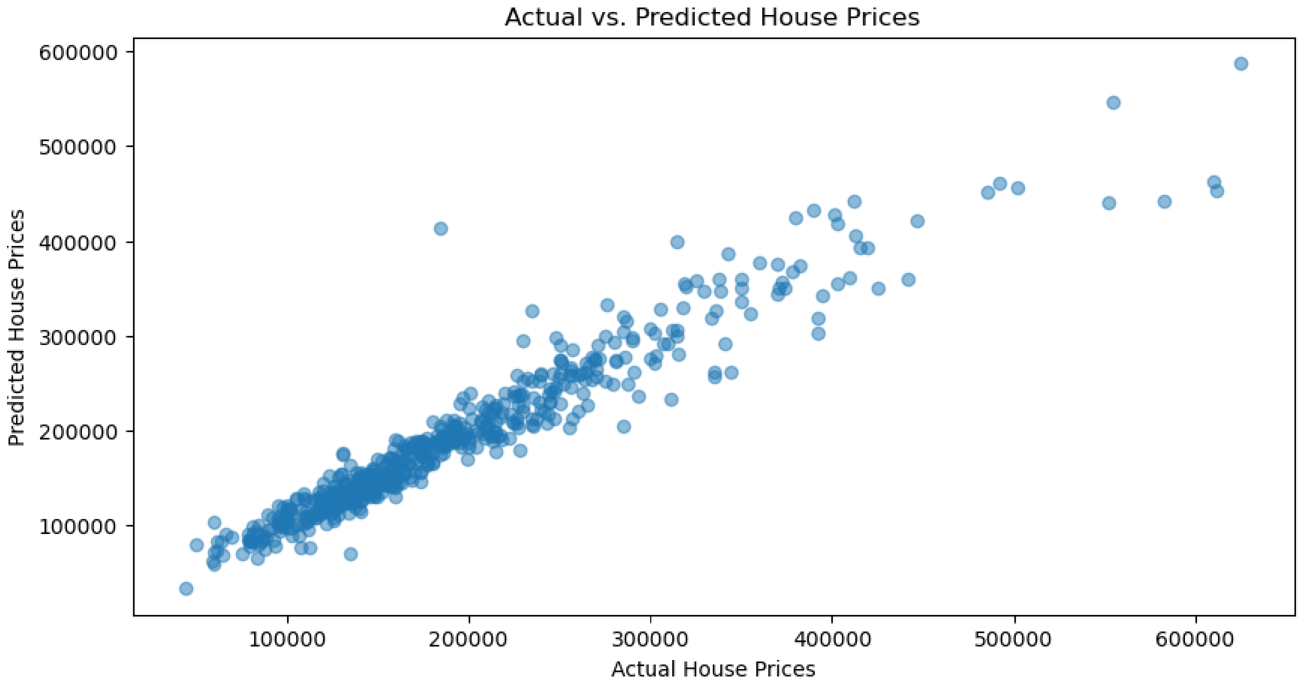
Training Process:

- Split dataset into training (80%) and testing (20%) sets.
- Applied 5-Fold Cross-Validation to assess model performance.
- Tuned hyperparameters using GridSearchCV.

Model Evaluation Metrics:

- Mean Absolute Error (MAE) = 15238.02
- Mean Squared Error (MSE) = 629375950.47
- R² Score = 0.9215

4. **Model Evaluation & Predictions:**



- Gradient Boosting Regressor performed best with an R² score of 0.8885
- Saved the best-performing model using Joblib for future predictions.
- Predicted house prices for new test samples.

Discussion of Results & Interpretations

Gradient Boosting Regressor performed best, achieving an R² score of 0.8885, making it the most accurate model. The analysis revealed that house prices in Ames are strongly influenced by structural quality, overall living area, and the presence of a garage. Correlation analysis confirmed that OverallQual and GrLivArea are the most impactful features, aligning with real estate valuation principles. While Random Forest Regression effectively handled nonlinearity and feature interactions, it did not surpass Gradient Boosting in performance. Linear Regression struggled due to the dataset's nonlinear relationships, highlighting the need for more advanced models. Gradient Boosting required more hyperparameter tuning and computational power but delivered superior accuracy. A key challenge was handling missing values. Some categorical variables with excessive missing data were removed, which may have led to minor information loss. Future improvements could involve advanced imputation techniques like KNN Imputation or deep learning-based models for price prediction. Overall, machine learning provides a strong framework for house price prediction. By integrating data preprocessing, feature selection, and model optimization, our approach enhances prediction accuracy—helping real estate professionals and investors make informed pricing decisions.

Conclusion

This project demonstrates the effectiveness of machine learning in real estate price prediction. By leveraging key housing features, our model provides reliable price estimates, aiding better decision-making for buyers, sellers, and investors. The project also highlights the importance of data preprocessing techniques such as feature engineering, handling missing values, and encoding categorical variables to enhance model performance. Future work could explore deep learning approaches or ensemble techniques to further improve predictive power.

Tools & Technologies

Python Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn,
Joblib Development Environment: Jupyter Notebook

References

Kaggle Dataset: Ames Housing

Machine Learning Mastery: ML in Python

Scikit-Learn Documentation: https://scikit-learn.org/

XGBoost Documentation: https://xgboost.readthedocs.io/en/stable/

OpenML Datasets: https://www.openml.org/