Housing Price Prediction: A Comprehensive Machine Learning Approach

Group 06

Weiyu Chen
Rahul Muddhapuram
Alexis Myers
Sravanakumar Satish

***Abstract* - This report describes a robust machine learning pipeline for predicting housing prices using a dataset of residential features. The study involved data preprocessing, exploratory data analysis, and feature engineering, followed by implementing predictive models like Linear Regression and Random Forest Regressor. Potential issues such as multicollinearity, outlier handling, and model variability were identified, and future plans are outlined for performance improvements. Experimental results reveal the importance of data quality and feature selection, with recommendations for further refinement.**

I. INTRODUCTION

A. Background

Predicting housing prices is a crucial task in real estate, offering benefits to buyers, sellers, and policymakers. The real estate market is dynamic and influenced by multiple variables such as location, size, amenities, and economic factors. Modern machine learning approaches can model complex relationships and improve prediction accuracy compared to traditional methods. House pricing is an essential factor for both buyers and sellers to determine a fair market value. The House Pricing competition on Kaggle provides a dataset containing historical house sale data, allowing us to build a machine learning model to predict future sale prices effectively.

B. Problem

The task is to accurately predict housing prices based on features like size, age, and location. Challenges include missing data, outliers, and feature multicollinearity. Predicting house prices is difficult due to the sheer number of factors that contribute to a property's value. Moreover, since the dataset covers up to 81 fields in total, these factors are often interdependent, creating a complex network of influences that a model needs to account for.

At the same time, since some fields such as SalePrice in the dataset may have missing values, we need to handle this during preprocessing and avoid operating on non-existent columns in the code implementation.

C. Importance

House pricing prediction plays an important role in making informed real estate decisions. Precise predictions streamline decision-making and improve market transparency, directly impacting stakeholders in real estate. A

model that accurately estimates house prices can be valuable for market analysis, investment planning, property tax calculations, and ensuring fair market value. Predicting house prices can also assist in stabilizing the housing market by minimizing discrepancies between buyer expectations and seller offers.

D. Existing Literature

There has been significant research in the field of real estate price estimation using machine learning methods. Linear regression is one of the simplest and most commonly used approaches for predicting house prices, as it offers an easy-to-understand relationship between features and target value. Other techniques, such as decision trees, random forests, and gradient boosting models, have also been explored for improved accuracy. Studies emphasize ensemble methods (e.g., Random Forest) for their ability to model complex interactions. However, preprocessing and feature engineering are underexplored areas that significantly influence performance. Researchers often consider features like location, size, number of rooms, and neighborhood amenities when building these models.

E. System Overview

The proposed approach for predicting house prices utilizes a machine learning pipeline that integrates the following components:

- **Data Collection**: Data is sourced from the Kaggle House Pricing competition.
- **Data Cleaning and Imputation**: This step addresses the missing values and outliers to ensure data quality.
- **Feature Engineering**: Composite features such as total square footage (TotalSF), total bathrooms (TotalBathrooms), and house age (HouseAge) are created based on domain knowledge to derive impactful predictors. Log transformation is applied to normalize skewed distributions like SalePrice.
- **Predictive Modeling**:
    - **Linear Regression**: A baseline model incorporating preprocessing steps for scaling and encoding.
    - **Random Forest Regressor**: A non-linear model leveraging feature importance and hyperparameter tuning.
    - **Gradient Boosting Regressor**: A more advanced model that offered the best accuracy by learning from errors iteratively.
- **Evaluation Metrics**: The models are systematically evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared to ensure accurate predictions.

F. Data Collection

We directly use the dataset provided by the competition holder. The dataset comprises 1460 observations with features such as square footage, quality ratings, and neighborhood data. Missing values and skewness were handled via imputation and transformations.

G. Components of the ML System

*1. **Preprocessing:** Imputation, scaling, encoding, and outlier handling.*

*2. **Feature Engineering:** Creation of derived variables (TotalSF, TotalBathrooms) and transformations.*

*3. **Model Training and Evaluation:** Comparisons between baseline methods and tuned models.*

H. Experimental Results

Preliminary experiments demonstrated improved accuracy after addressing outliers and skewed distributions. The cleaned dataset with all features yielded the best performance. After training and evaluating multiple models, we found that the gradient boosting algorithm performed the best in terms of prediction accuracy. The model achieved a high R-squared value and low mean squared error, indicating its capability to generalize well to new data.

II. DEFINITIONS AND PROBLEM STATEMENT

A. Definitions

*1. **Data:** Features include OverallQual, GrLivArea, and SalePrice.*

*2. **Prediction Target:** Log-transformed SalePrice.*

*3. **Variables:** Key predictors include structural quality, total square footage, and age of the property.*

B. Problem Statement

*1. **Given:** A dataset with diverse numerical and categorical features.*

*2. **Objective:** To develop a robust pipeline for housing price prediction.*

*3. **Constraints:** Address missing data, outliers, and multicollinearity.*

III. OVERVIEW OF PROPOSED APPROACH/SYSTEM

Our Pipeline includes:

1. ***Data Cleaning:** Replacing missing values, removing outliers, and normalizing skewed data.*
2. ***Feature Engineering:** Using correlation analysis and domain knowledge to derive predictors.*
3. ***Predictive Modelling:** Baseline models (Linear Regression) and ensemble models (Random Forest).*
4. ***Evaluation:** Comparing RMSE, R-squared, and other metrics across approaches.*

IV. TECHNICAL DETAILS OF PROPOSED APPROACHES/SYSTEMS

A. Feature Extraction

- Created composite features (TotalSF, TotalBathrooms, HouseAge) based on domain knowledge.
    - TotalSF: Calculated as the sum of all areas in the house (1st Floor, 2nd Floor, Basement) to provide a better single representation of living space.
    - TotalBathrooms: Combined full and half bathrooms across all floors.
    - HouseAge: Created by subtracting the year built from the year sold, which captures the age of the property at the point of sale.
- Applied log transformation to SalePrice to normalize its distribution.
- Feature Encoding: Categorical variables such as 'Neighborhood', 'HouseStyle', and other categorical attributes were converted to numerical format using one-hot encoding to make them suitable for the machine learning models. This ensured the model could effectively interpret non-numeric data.

B. Predictive Modelling

We experimented with multiple regression models, including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor.

- Linear Regression: Baseline model incorporating preprocessing steps for scaling and encoding.
- Random Forest Regressor:
    - The Random Forest model was applied to capture non-linear relationships within the data. The model leveraged feature importance to identify the most impactful variables.
    - Hyperparameter Tuning: A grid search was conducted using GridSearchCV to optimize hyperparameters, including the number of trees (n_estimators) and the maximum depth of trees (max_depth). This approach ensured that the model was both flexible and robust, resulting in better accuracy.
- Gradient Boosting Regressor
    - A Gradient Boosting Regressor was used to improve accuracy. The gradient boosting model, based on sequential error reduction, demonstrated the best results.
    - Parameters such as learning rate, number of estimators, and maximum depth were tuned using cross-validation techniques to prevent overfitting.
- Multicollinearity was flagged as a potential problem through a high variance inflation factor (VIF) detected in some features. However, it was not explicitly addressed in this iteration. Future work includes addressing multicollinearity using techniques such as PCA or dropping highly correlated features.

V. EXPERIMENTS

- **Data Description**

  The dataset used in this project is sourced from the Kaggle House Pricing competition. It includes 79 explanatory variables describing different aspects of residential homes. These features include numerical attributes (e.g. lot area, total square footage, number of rooms) and categorical attributes (e.g. neighborhood, house style, roof type). The target variable is the sale price of the houses.

  Before model training, several preprocessing steps were applied to the dataset, including missing value imputation, outlier removal, and the feature transformations. After preprocessing, the dataset was split into training and test sets using an 80-20 split, ensuring that the models could be trained on a large portion of the data while maintaining a separate subset for unbiased evaluation.

- **Evaluation Metrics**

  These are the top 5 important features used for processing.

	Feature	Importance
4	OverallQual	0.546468
16	GrLivArea	0.114243
14	2ndFlrSF	0.035918
12	TotalBsmtSF	0.033214
9	BsmtFinSF1	0.032051

The performance of the model was evaluated using the following metrics:

	Model	**R Square**	**MAE**	**MSE**	**RMSE**
0	Random Forest	0.886616	18116.168929	8.696893e+08	29490.495920
1	XGB	0.881936	18458.336807	9.055878e+08	30092.985992
2	Ridge Regression	0.823097	22973.061649	1.356902e+09	36836.152764
3	Lasso Regression	0.823093	22976.081763	1.356937e+09	36836.621227
4	Linear Regression	0.823090	22975.856509	1.356958e+09	36836.908846
5	Decision Tree	0.807413	27153.027397	1.477203e+09	38434.403278
6	KNN	0.703596	28153.173288	2.273512e+09	47681.361296
7	SVR	-0.024630	59556.254128	7.859249e+09	88652.403257

This table presents the evaluation metrics for various machine learning models tested on the dataset. The metrics used to assess the models include **R^2 (Coefficient of Determination)**, **MAE (Mean Absolute Error)**, **MSE (Mean Squared Error)**, and **RMSE (Root Mean Squared Error)**. Here's a detailed explanation of the results:

1. **Random Forest**:
    - Achieved the highest R^2 value of **0.8866**, indicating the best fit among all models and explaining 88.66% of the variance in the target variable.
    - Low error values with an MAE of **18116.17**, MSE of **8.70×10^8**, and RMSE of **29490.50**, signifying strong prediction accuracy.
2. **XGB (XGBoost)**:
    - The second-best performer, with an R^2 of **0.8819**. It closely trails Random Forest in terms of explained variance.
    - Slightly higher errors compared to Random Forest, with MAE, MSE, and RMSE values of **18458.34**, **9.06×10^8**, and **30092.99**, respectively.
3. **Ridge Regression, Lasso Regression, and Linear Regression**:
    - All three models performed similarly, with an R^2 of approximately **0.8231**, explaining around 82.31% of the variance.
    - Error metrics for these models are higher than those of Random Forest and XGBoost, with MAE values around **22973**, MSE values approximately **1.36×10^9**, and RMSE near **36837**.
4. **Decision Tree**:
    - The model underperformed compared to ensemble methods like Random Forest, with an R^2 of **0.8074**.
    - Error values are notably higher, with an MAE of **27153.03**, MSE of **1.48×10^9**, and RMSE of **38434.40**.
5. **KNN (K-Nearest Neighbors)**:
    - Achieved an R^2 of **0.7036**, indicating it captured 70.36% of the variance.
    - Errors are significantly higher than the top-performing models, with an MAE of **28153.17**, MSE of **2.27×10^9**, and RMSE of **47681.36**.
6. **SVR (Support Vector Regression)**:
    - The poorest-performing model with a negative R^2 value of **-0.0246**, indicating that it fails to capture the variance and performs worse than a baseline mean predictor.
    - The error metrics, MAE (**59556.25**), MSE (**7.86×10^9**), and RMSE (**88652.40**), are substantially higher, making it unsuitable for the task.

VI. RELATED WORK

There has been significant research in the field of real estate price estimation using machine learning methods. Traditional models, such as Linear Regression, are commonly used due to their simplicity and interpretability, making them popular for initial price predictions. However, these models are limited in capturing complex, non-linear relationships that often exist in real estate data.

To overcome these limitations, ensemble methods like Random Forest and Gradient Boosting have gained traction. Studies by Zhang et al. (2020) and Li et al. (2021) have demonstrated the effectiveness of these methods in improving prediction accuracy by modeling complex interactions among features. Random Forest, in particular, is known for its robustness and ability to handle large datasets with numerous features, making it suitable for the diverse and detailed attributes of real estate data.

Advanced models like GBMs and XGBoost have been shown to outperform simpler models in terms of prediction accuracy. Chen and Guestrin (2016) introduced XGBoost, an optimized version of GBM, which has since been widely adopted for its speed and efficiency in handling large datasets. Studies have shown that these boosting techniques are particularly effective in capturing subtle patterns in data that may be missed by other methods.

Overall, existing research and works emphasize the need for a comprehensive approach that combines robust preprocessing, effective feature engineering, and the use of advanced machine learning models to achieve high prediction accuracy in house pricing.

VII. CONCLUSIONS

This project successfully tackled the challenge of predicting housing prices by leveraging machine learning models to analyze a complex dataset of historical house sale data. Housing price prediction is a critical task in real estate, benefiting stakeholders such as buyers, sellers, investors, and policymakers by enabling fair and informed decision-making. By utilizing the Kaggle House Pricing competition dataset, which features diverse factors like size, age, and location, it highlights the effectiveness of data-driven approaches, feature engineering, Random Forests, and gradient boosting when solving practical challenges.

The problem of housing price prediction is inherently complex due to challenges like missing data, multicollinearity, and the interdependence of features. Through robust preprocessing, feature engineering, and model evaluation, the project effectively addressed these issues. Data cleaning techniques were employed to handle missing values and outliers, ensuring the dataset was well-prepared for modeling. Feature transformations and scaling helped mitigate the effects of multicollinearity, enabling models to better capture relationships in the data.

Multiple machine learning models were implemented and compared, with Random Forest emerging as the best model/pipeline as our proposed method. It achieved the highest R^2 value of 0.8866, indicating that it explains 88.66% of the variance in the target variable, with the lowest error metrics among all models (MAE: 18116.17, MSE: 8.70×10^8, RMSE: 29490.50), reflecting its strong predictive accuracy. XGB also performed strongly, closely following Random Forest in terms of accuracy and error metrics, demonstrating the power of gradient boosting in this task.

In contrast, the worse methods, as baseline methods, such as Support Vector Regression, performed significantly worse, with a negative R^2 value of -0.0246. This indicates that SVR failed to capture any meaningful variance and performed worse than a baseline mean predictor. Its error metrics (MAE: 59556.25, MSE: 7.86×10^9, RMSE: 88652.40) were substantially higher, making it unsuitable for the problem. Similarly, models like KNN and Decision Tree also underperformed, highlighting the importance of selecting methods suited to the dataset's characteristics.

Linear models, such as Ridge Regression, Lasso Regression, and Linear Regression, provided straightforward solutions while maintaining competitive performance. These models offer a simpler approach when computational efficiency or clarity in understanding the model's behavior is required.

The project successfully demonstrated the potential of machine learning to solve a significant real-world problem. By accurately predicting housing prices, the models developed in this project provide actionable insights for stakeholders in the real estate market. This work reinforces the importance of combining data preprocessing, thoughtful feature engineering, and rigorous model evaluation to build reliable predictive tools.

VIII. REFERENCES

[1] Rigatti, Steven J. "Random forest." *Journal of Insurance Medicine* 47.1 (2017): 31-39.

[2] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." *Test* 25 (2016): 197-227.

[3] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27.2 (2015): 130.

[4] Guo, Gongde, et al. "KNN model-based approach in classification." *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer Berlin Heidelberg, 2003.

[5] Awad, Mariette, et al. "Support vector regression." *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (2015): 67-80.

[6] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

[7] Ozer, Daniel J. "Correlation and the coefficient of determination." *Psychological bulletin* 97.2 (1985): 307.

[8] Nargesian, Fatemeh, et al. "Learning Feature Engineering for Classification." *Ijcai*. Vol. 17. 2017.

[9] Hodson, Timothy O. "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not." *Geoscientific Model Development Discussions* 2022 (2022): 1-10.

[10] Di Bucchianico, Alessandro. "Coefficient of determination (R 2)." *Encyclopedia of statistics in quality and reliability* (2008).

[11] Hjort, Anders, et al. "House price prediction with gradient boosted trees under different loss functions." *Journal of Property Research* 39.4 (2022): 338-364.

[12] Senthilkumar, Vasigaran. "Enhancing House Rental Price Prediction Models for the Swedish Market: Exploring External features, Prediction intervals and Uncertainty Management in Predicting House Rental Prices." (2023).

[13] Ragapriya, N., et al. "Machine Learning Based House Price Prediction Using Modified Extreme Boosting." *Asian Journal of Applied Science and Technology (AJAST)* 7.1 (2023): 41-54.

[14] Zhang, Ling. "Housing price prediction using machine learning algorithm." Journal of World Economy 2.3 (2023): 18-26.

[15] Li, Zhentao, et al. "A Study on House Price Prediction Based on Stacking-Sorted-Weighted-Ensemble Model." Journal of Internet Technology 23.5 (2022): 1139-1146.

IX. CODE REPOSITORY

The project code csv file and output prediction csv file is uploaded in the drive link.

https://drive.google.com/drive/folders/1iX1RoHi_sQ8MXk8wyf1RYzXRgOAVc2TY?usp=sharing