PREDICTING HOME PRICE IN AMES, IOWA USING ADVANCED REGRESSION TECHNIQUES

Anbu Suriya Kumar
Graduate Student in Data Science, DePaul University, Chicago

ABSTRACT:

This project presents a comprehensive analysis of house price prediction. It uses information about houses that have already been sold in Ames, Iowa. As the real estate market gets more and more complicated, being able to predict house prices accurately is becoming increasingly important. This is helpful for many people involved in buying and selling houses, such as buyers, sellers, and real estate agents. The dataset comprises a rich collection of structural, spatial, and neighborhood attributes, providing a robust foundation for modeling. The study utilizes rigorous data preprocessing techniques including handling missing values, encoding categorical variables, and scaling features. Feature selection methods, like Mutual info regression, wrapper, univariate selection, are employed to identify the most impactful features for modelling. Hyperparameter tuning optimizes the parameters of the regression models, enhancing their ability to predict. Additionally, the XGBoost algorithm is implemented to further improve prediction accuracy. Evaluation metrics like Root Mean Squared Error (RMSE) and explained variance assess model performance. The findings reveal effective strategies for accurate house price prediction, offering valuable insights for real estate stakeholders and data scientists.

1.1 INTRODUCTION:

In the realm of real estate, one of the major challenges faced by homebuyers is the lack of transparency in pricing. Sellers often set prices based on subjective factors such as personal preferences, emotional attachment to the property, or even speculative market trends. This lack of clarity can lead to a gap between a property's perceived value and its true market worth. Predictive modeling offers a data-driven solution to this challenge by analyzing past sales data, property features, and market trends, predictive models can estimate a property's fair market value. This empowers homebuyers to make informed decisions and negotiate effectively, ensuring they pay a fair price. Moreover, predictive modeling in house pricing extends beyond individual transactions and has broad applications in various aspects of real estate, such as assessing ROI for investors, forecasting housing demand for government agencies, and informing urban development strategies. However, Traditional real estate transactions often involve intermediaries like agents, increasing costs for both buyers and sellers. This has led to growing interest in alternative approaches that use predictive modeling and technology to streamline the process. This shift towards a data-driven and technology-enabled approach has the potential to transform the real estate industry and improve accessibility and affordability for homebuyers.

This project explores and analyzes various predictive modeling techniques for house price prediction using the Ames housing dataset. Building on existing research using regression models, this project will delve into the effectiveness of advanced machine learning algorithms like decision trees, random forests, gradient boosting, and XGBoost. Additionally, it will employ techniques like grid search for hyperparameter tuning and feature selection methods to improve model performance. The project will compare the predictive accuracy of these models with and without feature selection to assess the impact of feature engineering. Finally, it will investigate the effects of normalizing the target variable and explore the impact of encoding methods on model performance.

1.2 DATASET:

The Ames housing dataset is a rich collection of housing-related data from the city of Ames, Iowa. It contains a wide range of information about residential properties, including both numerical (38 variables) and categorical (43 variables) features. The dataset includes attributes such as property size, number of bedrooms and bathrooms, overall quality and condition ratings, zoning classifications, street types, and neighborhood details. Additionally, it provides the sale prices of properties, which is the target variable for our predictive models.

2. RELATED WORK

Kok et al. [1] investigated the performance of various machine learning algorithms for real estate price appraisals. They analyzed a dataset of 84,305 observations from California, Florida, and Texas between 2011 and 2016. The study compared ordinary least squares regression (OLS) with Random Forest (RF), Gradient Boosting Regression (GBR), and XGBoost (XGBM) techniques. Their findings suggest that XGBM generally outperformed the other algorithms for real estate price prediction.

In another relevant study by Luo et al. (2022), the authors propose a novel ensemble learning framework specifically designed for house price prediction [2]. This framework acknowledges the importance of spatial dependence and heterogeneity, factors often overlooked in traditional models. Their approach incorporates geographically weighted regression (GWR) to capture these spatial variations and utilizes a combination of machine learning algorithms, including XGBoost and Support Vector Machines (SVM), to enhance the model's generalizability. The proposed framework demonstrates superior performance compared to existing methods, particularly in areas with significant spatial variations in house prices.

Ahsan et al. (2021) investigated the impact of data scaling methods on the performance of machine learning algorithms [3]. Their study focused on a dataset containing information about heart disease patients. They evaluated the performance of eleven machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, alongside six different data scaling methods like normalization and standardization. The results indicated that the choice of data scaling method can significantly influence the performance of machine learning models. The study found that the best performing algorithm-data scaling method combination depended on the specific dataset and machine learning task.

Bentéjac et al. (2021) conducted a comparative analysis of gradient boosting algorithms, which are a powerful class of machine learning algorithms used for regression and classification tasks [4]. Their study focused on three recently proposed gradient boosting algorithms: XGBoost, LightGBM, and CatBoost. They evaluated these algorithms on a variety of benchmark datasets and compared their performance in terms of training speed, generalization accuracy, and hyperparameter sensitivity. The findings revealed that CatBoost achieved the best overall results in terms of generalization accuracy, while LightGBM emerged as the fastest training algorithm.

However, the differences in performance between the algorithms were found to be relatively small.

Estevez et al. (2009) proposed a feature selection method based on Normalized Mutual Information (NMI) in their research published on Neural Networks [5]. Feature selection is a crucial step in machine learning tasks, aiming to identify the most relevant features that contribute to the model's performance. Traditional feature selection methods might not always capture the complex relationships between features and the target variable. The proposed NMI-based method offers a data-driven approach for feature selection. NMI quantifies the mutual dependence between features and the target variable, allowing the selection of features that hold the most informative relationship for prediction. This approach can be particularly beneficial in scenarios where features exhibit complex dependencies.

4. METHODOLOGY:

4.1.1 Feature Engineering:

As part of the feature engineering process, a new feature named TotalSF was created by summing the TotalBsmtSF, 1stFlrSF, and 2ndFlrSF variables. This composite feature provides a comprehensive measure of the total square footage of each house. Additionally, certain numerical variables that were actually categorical in nature, such as OverAllQual, MSSubClass, and OverAllCond, were appropriately transformed into categorical variables based on their values.

In this study, used label encoding to convert categorical variables into numerical format. This allows to incorporate these variables into the predictive modeling process effectively. By transforming categorical data into a format that machine learning algorithms can understand, it leverages the information encoded in these variables to make accurate predictions about house prices in Ames, Iowa.

4.1.2 Exploratory Data Analysis (EDA):

The EDA phase commenced by examining the distribution of the target variable, SalePrice, through a distribution plot [Fig 4.4]. The plot revealed a skewed distribution with more frequent occurrences of lower sale prices on the left side, accompanied by a tail extending to the right indicating a presence of expensive outliers. A correlation matrix [Fig 4.5] was employed to identify numerical features highly correlated with the target variable. Noteworthy features found to be strongly correlated included TotalSF, GrLivArea, GarageCars, YearBuilt, YearRemodAdd, and TotRmsAbvGrd.To address outliers in numerical variables, box plots were utilized along with the z-score method, which facilitated the identification and handling of outliers by calculating the standard score of each data point. The relationship between highly correlated features and the target variable was further investigated using scatter plots, revealing a linear relationship between TotalSF and GrLivArea.

Conducting an ANOVA test for categorical variables with the target variable, SalePrice, provided insights into their impact on house prices. Significant variables identified through this analysis included HeatingQC, Neighborhood, BsmtExposure, and BsmtFinType1. Subsequent analysis of the relationship between HeatingQC and SalePrice [Fig 4.3] indicated that houses with excellent heating conditions tended to command higher prices, echoing similar trends observed for other categorical variables. Visualizing the sale prices of houses in each neighborhood [Fig 4.4] enabled the identification of the best-performing (Mitchell) and worst-performing (Edwards) neighborhoods based on sale prices. Further exploration of the HeatingQC and Bsmt Quality (BsmtFinType1) features within these neighborhoods revealed houses in Mitchell exhibited better basement quality and heating conditions compared to Edwards, correlating with the differences in sale prices between the two neighborhoods.

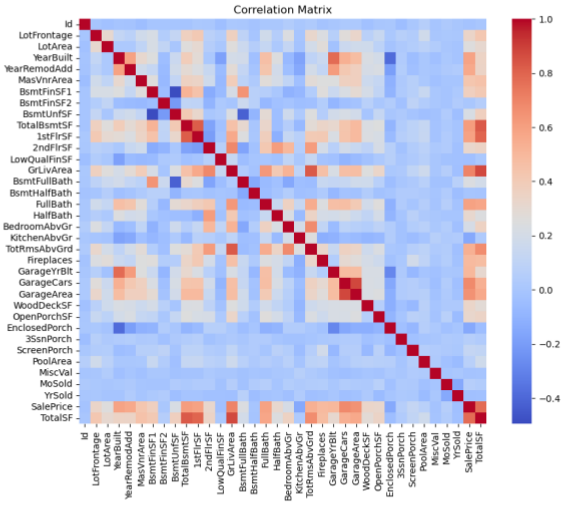


Fig. 4.1

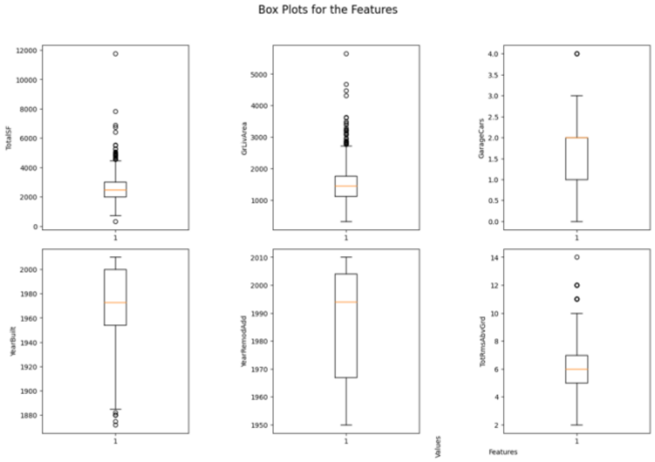


Fig. 4.2

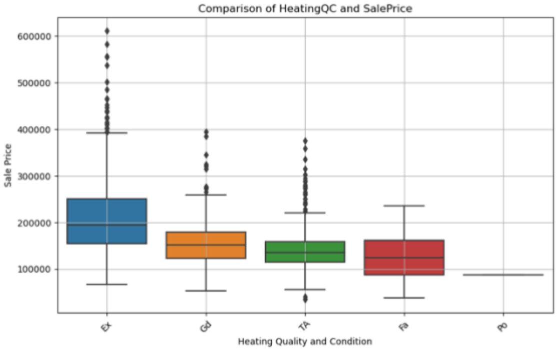


Fig. 4.3

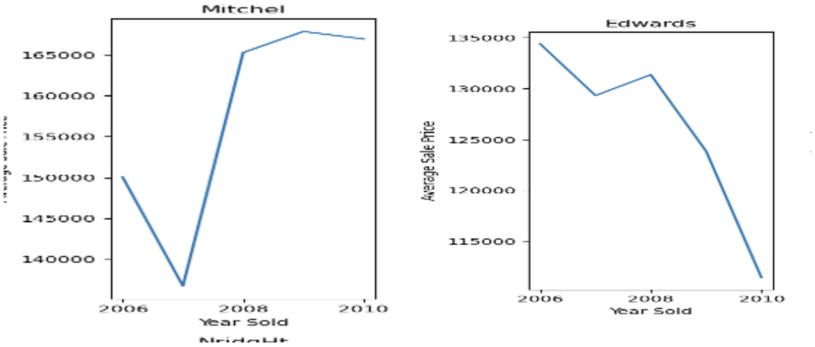
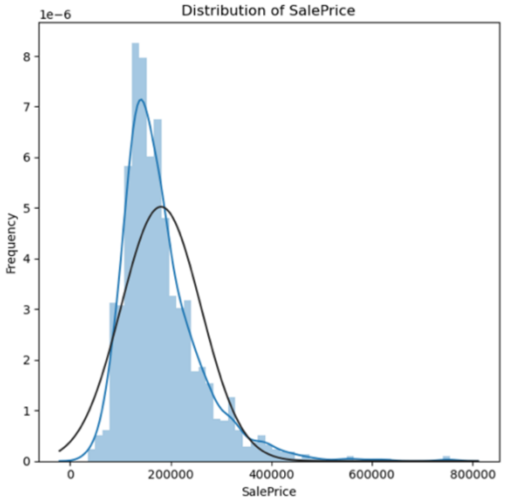


Fig 4.4



The comprehensive data exploration undertaken in this phase provided valuable insights into the factors influencing house prices in the Ames dataset. It underscored the significance of features such as total square footage, neighborhood characteristics, and heating quality in determining house prices. Additionally, it highlighted the importance of rigorous data preprocessing and exploratory analysis in preparing the dataset for subsequent predictive modeling tasks.

Fig 4.5

4.2 Model building:

In this phase of the project, various regression models were employed to predict house prices using the features identified during the exploratory data analysis (EDA) and features identified in feature engineering using Wrapper selection, Univariate selection, and Mutual info regression. The models were evaluated based on their ability to accurately predict house prices, measured using metrics such as Root Mean Squared Error (RMSE) and explained variance (Expl Var). The models considered include Decision Trees, Random Forest, Gradient Boosting, and XGBoost.

Initially, I conducted a test run of my model with default parameters for the non-logarithmic transformed target variable to assess the results. Subsequently, I reran the models after logarithmically transforming the target variable and interpreted the results. It became evident that applying the transformation to my data yielded better results. Therefore, I proceeded with the analysis of selected features for all the models using the transformed data.

Decision Trees:

Decision Trees are a type of supervised learning algorithm used for classification and regression tasks. In the context of house price prediction, Decision Trees split the dataset into subsets based on the value of features, aiming to minimize the RMSE. For this study, I have used the criterion parameter as "squared_error", n_estimators =100 and other default params to determine the results.

Random Forest:

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree in the Random Forest is trained on a subset of the data, and the final prediction is made by averaging the predictions of all trees. In this study, the number of estimators was varied to assess its impact on model performance and have selected the best performing estimator to be 100 with criterion "squared error".

Gradient Boosting:

Gradient Boosting is another ensemble learning technique that builds decision trees sequentially, with each tree attempting to correct the errors made by the previous tree. Gradient Boosting optimizes a differentiable loss function, such as mean squared error (MSE) or huber loss, to minimize prediction errors [1]. The learning rate, number of estimators, and maximum depth of trees were tuned to improve model performance.

XGBoost:

XGBoost [1] [2] [3] stands out as an optimized implementation of Gradient Boosting, offering enhanced performance and efficiency. It incorporates a more regularized model structure to effectively control overfitting and integrates additional features such as parallelization and tree pruning. During the training process, XGBoost was fine-tuned using various hyperparameters, including the learning rate, maximum depth, and number of estimators, aiming to achieve the optimal configuration for house price prediction. Hyperparameter tuning was conducted using grid search, with the parameter grid defined as follows: { 'n_estimators': [100, 200, 300], 'learning_rate': [0.05, 0.1, 0.2], 'max_depth': [3, 4, 5], 'min_samples_split': [2, 3, 4], 'loss': ['squared_error', 'absolute_error', 'quantile', 'huber'], 'random_state': [42] }.

Wrapper Method:

It builds and evaluates different models, each using a unique combination of features as its toolbox. By comparing the performance of these models (often using metrics like RMSE), the wrapper method identifies the subset of features that leads to the most accurate predictions. This approach is like having a team meeting and collaboratively choosing the best tools for the task. While it might require some extra time, the wrapper method often uncovers powerful combinations of features that can significantly improve model performance.

Selected features: '1stFlrSF', '2ndFlrSF', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'CentralAir', 'GrLivArea', 'LowQualFinSF', 'TotalBsmtSF', 'TotalSF'

Univariate Selection:

It analyzes each feature individually, assessing its relationship with the target variable. This is like having a wise expert who examines each tool and assigns it a score based on its perceived importance for the job at hand. Features with the highest scores are then chosen for the final model. While it might not capture complex interdependencies between features, univariate selection can be a fast and efficient way to identify a good starting point for feature selection.

Selected Features: 1stFlrSF', 'ExterQual', 'FullBath', 'GarageArea', 'GarageCars', 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'TotalSF', 'YearBuilt'

Mutual Information Regression:

It calculates a score for each feature, reflecting how much information it shares with the target variable. Features with high scores are considered more informative and are selected for the model [5].

Selected Features: 'ExterQual', 'GarageArea', 'GarageCars', 'GrLivArea', 'KitchenQual', 'Neighborhood', 'OverallQual', 'TotalBsmtSF', 'TotalSF', 'YearBuilt'

Selected features based on EDA: TotalSF', 'GrLiveArea', 'GarageCars', 'Yearbuilt', 'YearRemodAdd', 'TotRmsAbvGrd', 'HeatingQC', 'Neighborhood', 'BsmtFinType1'

5.1 Results and Discussion:

After training each model with the selected features, the performance of each model was evaluated using RMSE and Expl Var metrics. The results are summarized in the table 1.1, table has best RMSE and Expl Var score produced by each model cross various feature selection methods.

The performance of each model varied based on the selected features and hyperparameters. Random Forest and Gradient Boosting achieved the lowest RMSE values, indicating superior predictive accuracy compared to Decision Trees and XGBoost. However, Gradient Boosting with the feature selection method based on personal analysis yielded the highest explained variance, indicating its ability to capture more variance in house prices.

Method	Feature selection method with best results	RMSE	Expl Var
Decision Tree	Wrapper method	0.19	0.77
Random Forest	Wrapper method	0.13	0.89
Gradient Boosting	My Selection	0.12	0.90
XGB	Wrapper method	0.17	0.83

Table 1.2

From the table 1.1, we could see the best performing model was Gradient boosting. So next I have applied hyperparameter tuning with grid search to find the best hyperparameters for each feature selection method. The results are summarized in Table 1.2

Feature selection method	RMSE	Expl Var	Best Params
All features	0.11	0.81	learning_rate: 0.1, loss: huber, max_depth: 3, min_samples_split: 4, n_estimators: 300, random_state: 42
Wrapper method	0.12	0.83	learning_rate: 0.5, loss: huber, max_depth: 4, min_samples_split: 4, n_estimators: 100, random_state: 42
Univariate	0.12	0.81	learning_rate: 0.1, loss: huber, max_depth: 3, min_samples_split: 4, n_estimators: 200, random_state: 42
Mutual Regression	0.13	0.85	learning_rate: 0.1, loss: huber, max_depth: 3, min_samples_split: 4, n_estimators: 300, random_state: 42
My Selection	0.11	0.82	learning_rate: 0.1, loss: huber, max_depth: 3, min_samples_split: 4, n_estimators: 300, random_state: 42

Table 1.2

Hyperparameter tuning led to further improvements in RMSE and explained variance for all feature selection methods. The optimal hyperparameters varied slightly depending on the selected features, highlighting the importance of fine-tuning model parameters to achieve the best predictive performance.

5.2 Conclusion:

Based on the evaluation of each model using RMSE and Expl Var metrics, the performance varied depending on the selected features and hyperparameters. Notably, Random Forest and Gradient Boosting exhibited the lowest RMSE values, indicating superior predictive accuracy compared to Decision Trees and XGBoost. However, it's essential to consider the explained variance metric as well, where Gradient Boosting with the feature selection method based on personal analysis demonstrated the highest explained variance, suggesting its capability to capture more variance in house prices.

Further analysis revealed that Gradient Boosting emerged as the best-performing model overall. Consequently, hyperparameter tuning was applied using grid search to identify the optimal hyperparameters for each feature selection method. The results, summarized in Table 1.2, showcased enhancements in both RMSE and explained variance across all feature selection methods. The slight variations in optimal hyperparameters underscored the importance of fine-tuning model parameters to achieve optimal predictive performance.

Future iterations of this study could explore the incorporation of additional features, such as an amenity score for the neighborhood derived from external data sources. Factors like proximity to grocery stores, schools, hospitals, parks, and other amenities, as well as distance from downtown, could provide valuable insights into property valuation. By integrating such supplementary features into our analysis, we can gain a more comprehensive understanding of the factors influencing house prices and further refine our predictive models.

REFERENCES:

[1] Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). "Big Data in Real Estate? From Manual Appraisal to Automated Valuation." *Special Real Estate Issue 2017*

[2]Anders Hjort, Johan Pensar, Ida Scheel & Dag Einar Sommervoll (2022) House price prediction with gradient boosted trees under different loss functions, Journal of Property Research, 39:4, 338-364, DOI: 10.1080/09599916.2022.2070525

[3] Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*. 2021; 9(3):52. DOI: https://doi.org/10.3390/technologies9030052

[4] Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54, 1937–1967 (2021). DOI: https://doi.org/10.1007/s10462-020-09896-5

[5] P. A. Estevez, M. Tesmer, C. A. Perez and J. M. Zurada, "Normalized Mutual Information Feature Selection," in IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189-201, Feb. 2009, DOI: 10.1109/TNN.2008.2005601