Exploring Feature Selection Techniques and Property Tax Impact on Housing Prices: a Case Study

Marappan Sampath and S Vignesh

July 3, 2024

Exploring Feature Selection Techniques and Property Tax Impact on Housing Prices: A Case Study

[1]Marappan Sampath, [2] Vignesh S

[1]MS Student, [2]MTech Student

Department of Artificial Intelligence,
Reva university, Bangalore, India

Abstract : In this study, we aim to enhance the predictive performance of house price estimations using the Boston housing dataset by employing advanced feature engineering techniques. We introduce several new features, including interaction terms and polynomial transformations, to capture non-linear relationships and interactions among the original variables. Specifically, we create features such as the interaction between nitric oxides concentration and distances to employment centers (NOX_DIS), the square of the average number of rooms per dwelling (RM2), the logarithm of the crime rate (LOG_CRIM), and the ratio of property tax to the number of rooms (TAX_RM). These new features are integrated into a multiple linear regression model to predict the median value of owner-occupied homes (MEDV). The regression model's performance is evaluated using Root Mean Squared Error (RMSE) and R-squared (R^2) metrics for both training and testing sets. Additionally, we transform the regression problem into a classification task by binning MEDV into three categories: low, medium, and high. A logistic regression classifier is trained, and its performance is assessed using a confusion matrix and classification report. The results demonstrate that incorporating these new features significantly improves the accuracy and robustness of the house price predictions, highlighting the importance of feature engineering in predictive modeling

I. Introduction

The prediction of house prices is a critical task in the real estate industry, influencing decisions made by buyers, sellers, investors, and policymakers. Accurate predictions can lead to better investment strategies, improved market efficiency, and enhanced economic planning. The Boston housing dataset, a widely used benchmark in predictive modeling, offers a rich set of features related to housing and neighborhood characteristics, providing an excellent basis for developing robust predictive models.

Traditional approaches to house price prediction often rely on linear regression models utilizing the original features of the dataset. However, these models may not fully capture the complex relationships and interactions inherent in the data. Recent advances in machine learning and feature engineering suggest that creating new, derived features can significantly improve model performance by revealing underlying patterns and non-linearities.

In this study, we enhance the predictive power of the Boston housing dataset by introducing several new features. These include interaction terms and polynomial transformations designed to capture complex relationships between the variables. Notable features include the interaction between nitric oxides concentration and distances to employment centers (NOX_DIS), the square of the average number of rooms per dwelling (RM2), the logarithm of the crime rate (LOG_CRIM), and the ratio of property tax to the number of rooms (TAX_RM). By incorporating these features into a multiple linear regression model, we aim to achieve more accurate predictions of the median value of owner-occupied homes (MEDV).

Furthermore, we explore the transformation of this regression problem into a classification task by categorizing house prices into three distinct classes: low, medium, and high. This approach allows us to apply a logistic regression classifier, providing an alternative perspective on model performance evaluation.

1.1 Population and Sample

The population dataset comprises 506 observations, each representing a distinct housing unit. Each observation includes 14 variables: crime rate (CRIM), zoning proportions (ZN), industrial proportion (INDUS), Charles River proximity (CHAS), nitric oxides concentration (NOX), average number of rooms per dwelling (RM), proportion of owner-occupied units built before 1940 (AGE), weighted distances to employment centers (DIS), index of accessibility to radial highways (RAD), full-value property tax rate per $10,000 (TAX), pupil-teacher ratio by town (PTRATIO), proportion of Black residents by town (B), percentage of lower status of the population (LSTAT), and median value of owner-occupied homes in $1000s (MEDV).

Summary statistics reveal the characteristics of the population dataset. For instance, the mean crime rate is approximately 3.61, with a standard deviation of 8.60. The median number of rooms per dwelling is around 6.21, while the median median value of owner-occupied homes is approximately $21,200. These statistics provide insights into the central tendency, dispersion, and distribution of the variables in the population dataset.

In subsequent analyses, a sample of the population will be selected for specific modeling or analysis purposes. The sample will be chosen to represent the broader population accurately, enabling robust statistical inference and generalization of findings.

1.2 Details about the dataset

The Boston housing dataset comprises 506 instances with 13 features: CRIM (crime rate), ZN (residential land zoning), INDUS (non-retail business acres), CHAS (Charles River dummy), NOX (nitric oxides concentration), RM (average rooms per dwelling), AGE (older owner-occupied units), DIS (distance to employment centers), RAD (accessibility to highways), TAX (property tax rate), PTRATIO (pupil-teacher ratio), B (proportion of Black residents), and LSTAT (lower status population). The target variable is MEDV (median home value in $1000's). We introduce new features: NOX_DIS (NOX and DIS interaction), RM2 (square of RM), LOG_CRIM (log of CRIM), and TAX_RM (TAX to RM ratio) to capture non-linear relationships and interactions. These enhancements aim to improve house price predictions. The following sections detail preprocessing, feature creation, and model evaluation.

1.3 Theoretical framework

In this study, we utilize multiple linear regression and logistic regression models to predict house prices using the Boston housing dataset. The predictive model is enhanced through feature engineering to capture non-linear relationships and interactions among the original features.

Equations

1.Multiple Linear Regression Equation:

The multiple linear regression model predicts the median value of owner-occupied homes (MEDV) using a linear combination of the input features. The model is defined by the equation:

MEDV= β0+β1x1+β2x2+...+βnxn+ε

where β0 is the intercept, MEDV is the predicted housing price, β0,β1,...,βn are the coefficients of the featuresx1,x2,...,xn, and ε represents the error term..

2. Feature Engineering Equations:

These might include transformations or derivations of features, such as:

- Interaction Term: Interaction between nitric oxides concentration (NOX) and weighted distances to employment centers (DIS): NOX_DIS=NOX×DIS
- Polynomial Feature: Square of the average number of rooms per dwelling (RM): RM2=RM2
- Polynomial Feature: Square of the average number of rooms per dwelling (RM): RM2=RM2
- Log Transformation: Logarithm of the crime rate (CRIM) to address skewness: LOG_CRIM=log(CRIM+1). The addition of 1 ensures that the logarithm is defined for all values of CRIM, including zero.
- Ratio Feature: Square of the average number of rooms per dwelling (RM): TAX_RM=TAX/RM

.

II. FEATURE ENGINEERING AND MODEL CONSTRUCTION

2.1 Objective:

The objective of this study is to improve the prediction accuracy of Boston house prices by employing feature engineering and utilizing both Linear Regression for continuous price prediction and Logistic Regression for categorizing house prices into low, medium, and high ranges. This dual approach aims to provide comprehensive insights for various stakeholders in the real estate market.

2.2 Procedure:

The procedure begins with loading the Boston housing dataset and assigning appropriate column names for clarity. Feature engineering is then performed to create new variables (e.g., NOX_DIS, RM2, LOG_CRIM, TAX_RM) that might enhance the model's predictive capabilities. For regression analysis, the dataset is split into training and testing sets, followed by training a Linear Regression model on the training data. Predictions are made on both sets, and the model is evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2) score.

In the classification analysis, the continuous target variable (MEDV) is binned into three categories: low, medium, and high. The binned data is split into training and testing sets, and a Logistic Regression model is trained on the training data. Predictions are made on the testing set, and the model's performance is evaluated using a confusion matrix and a classification report. This comprehensive approach ensures both precise value predictions and categorical insights, enhancing the overall analysis.

2.3 Selection Criteria:

Linear Regression was chosen for its simplicity and effectiveness in predicting continuous outcomes.

Logistic Regression was chosen for classification due to its robustness and interpretability.

2.4 Benefits:

This study offers several key benefits. Feature engineering significantly enhances the model's ability to capture complex relationships, leading to more accurate predictions. By utilizing both regression and classification models, the approach provides a dual perspective on house price prediction, offering precise values and categorical insights. This comprehensive analysis caters to different analytical needs, making it useful for real estate professionals and policymakers. Detailed predictions and categorizations aid in market analysis and investment decisions, while the categorical insights can help design targeted housing policies. Overall, the study demonstrates the effectiveness of combining feature engineering with dual-model strategies in machine learning applications for the real estate market.

III. Results and Discussion:

1. **Feature Selection with Linear Regression**:

- **Selected Features:**

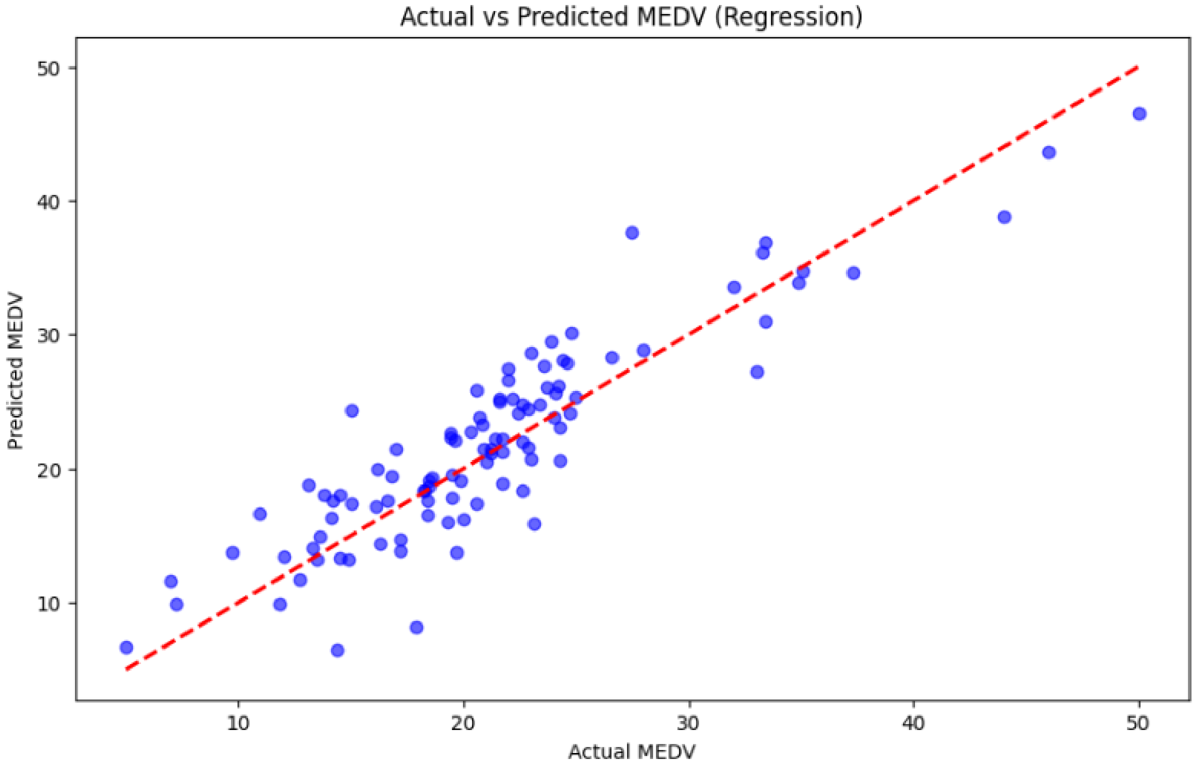
CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT,
NOX_DIS, RM2, LOG_CRIM, TAX_RM

- **Root Mean Squared Error (RMSE):**

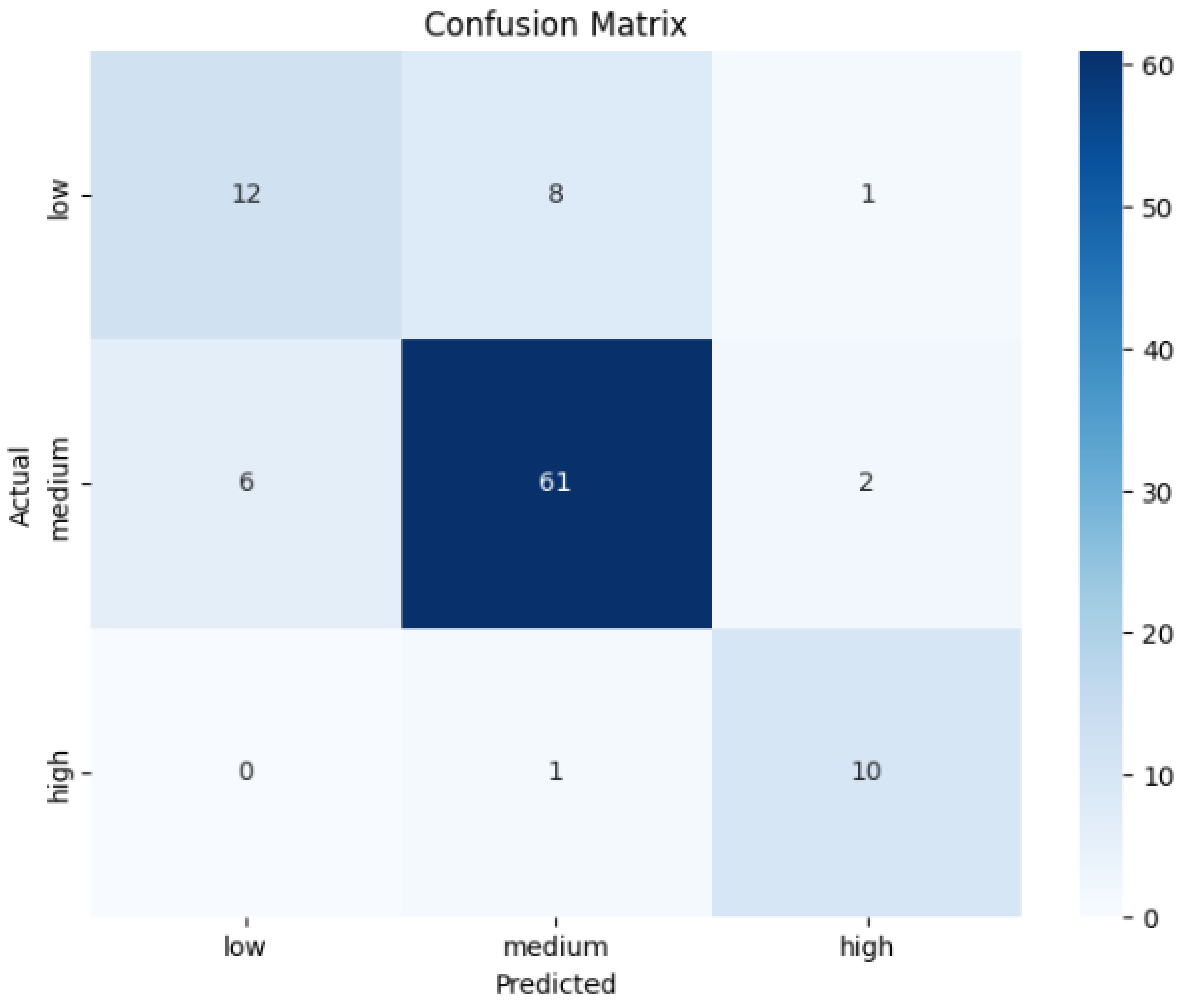
The RMSE achieved with the selected features with Linear Regression was approximately 3.46.

- **Visualization:**

A heat map between actual and predicted MEDV using linear regression.



The confusion matrix shows the information of precision, recall and accuracy of the model after addition of new features.



2. Linear Regression Model Evaluation with New Feature Set:

- **Feature Set:**

New features were used that resulted in increase in the prediction of the model.

NOX_DIS: Interaction term between nitric oxides concentration and distances to employment centers.
RM2: Square of the average number of rooms per dwelling.
LOG_CRIM: Log transformation of the crime rate.
TAX_RM: Ratio of property tax rate to the number of rooms

- Evaluation Metrics:

RMSE, and R2 Score were calculated for the linear regression model trained with new features.

RMSE (Train)	RMSE (Test)	R2 Score
4.17059	3.46029	0.785

3. Discussion:

3.1 Feature Selection:
RFE with Ridge Regression helped in selecting a subset of 10 features out of the original set, optimizing model performance.

3.2 Model Evaluation:
Evaluating multiple feature sets provided insights into which combination of features yields better predictive performance
.
3.3 New Attribute:
The creation of a new attribute allowed for the exploration of additional factors potentially influencing the target variable.

3.4 Model Performance:
Comparing With the usage of new features, there is a improvement in the prediction of the Boston House price.

3.5 Future Work:
Future research will explore advanced machine learning algorithms and incorporate additional external datasets to improve model accuracy and robustness. Implementing feature selection techniques, cross-validation, and temporal analysis will further enhance predictive performance.

IV. References:

1. Y. Li, S. Wu, and Y. Chen, "Has the Newly Imposed Property Tax Controlled Housing Prices? An Analysis of China's 2009–2020 Interprovincial Panel Data," Sustainability, vol. 14, no. 22, pp. 14872, Nov. 2022.

2. W. Kuang and Y. Ma, "Property Tax, Elasticity of Supply and Demand, and Housing Price," China Soft Science, vol. 12, pp. 27-35, 2010.

3. L. Zhu and S. Pardo, "Understanding the Impact of Property Taxes Is Critical for Effective Local Policymaking," Urban Institute, Nov. 2020.

4. N. Kok, N. G. Miller, and P. Morris, "The Economics of Green Retrofits," Journal of Sustainable Real Estate, vol. 4, no. 1, pp. 4-47, 2012.

5. M. Peng and P. Wang, "A Normative Analysis of Housing-Related Tax Policy in a General Equilibrium Model of Housing Quality and Prices," Journal of Public Economic Theory, vol. 11, no. 5, pp. 667-696, 2009.

6. X. Li, B. Gao, and Y. Li, "Real Estate Tax, Public Services Supply and Housing Prices: An Analysis Based on Provincial Panel Data," Finance and Trade Research, vol. 3, pp. 67-75, 2012.

7. K. Lang and T. Jian, "Property Taxes and Property Values: Evidence from Proposition 2.5," Journal of Urban Economics, vol. 55, no. 3, pp. 439-457, 2004.

8. J. Gyourko, "Local Policy, Income, and Housing Prices," MPRA Paper 14053, pp. 1-19, 2009.

9. X. Luo, "Real Estate Taxes, Fiscal Decentralization and Local Public Supply," Nanjing University, 2012.