JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

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

[1]Vijay Pawar H G, [2] Sumit Prakash Dubey

[1]Mtech Student, [2]Mtech Student

Department of Artificial Intelligence,
Reva university, Bangalore, India

Abstract : This study explores feature selection and regression modeling strategies for predicting median house values using a housing dataset. Recursive Feature Elimination (RFE) with Ridge Regression is employed to identify the most influential features. Performance evaluation metrics such as Mean Squared Error (MSE) and R-squared (R2) are utilized to assess model accuracy. Furthermore, the correlation between property tax burden and house value is examined. The findings offer valuable insights into housing market dynamics and provide practical implications for real estate stakeholders and researchers aiming to enhance prediction accuracy and understand key factors influencing house prices.

I. INTRODUCTION

The housing market stands as one of the most critical sectors influencing economic stability and societal well-being. Understanding the intricate dynamics of housing prices is not only essential for homeowners and real estate investors but also for policymakers and urban planners. In this context, predictive modeling emerges as a potent tool for forecasting housing prices, aiding in decision-making processes, and identifying influential factors shaping market trends.

This paper delves into the development of a predictive model for housing prices utilizing machine learning techniques. Leveraging a dataset encompassing various socio-economic and environmental attributes, we aim to construct a robust model capable of accurately predicting median home values. Through this endeavor, we seek to uncover the underlying relationships between housing prices and a plethora of factors, including crime rates, zoning regulations, air quality indices, and more.

The methodology employed encompasses several key steps. Firstly, we conduct exploratory data analysis to gain insights into the distribution and interplay of different features. Visualization techniques such as histograms and box plots offer a comprehensive overview of the dataset, facilitating feature selection and model development.

Subsequently, we delve into preprocessing tasks, including handling missing values and standardizing features to ensure model stability and performance. Utilizing techniques such as imputation and feature scaling, we prepare the dataset for further analysis and modeling.

Feature selection emerges as a crucial aspect of model development, aiming to identify the most influential variables driving housing prices. Through techniques like Recursive Feature Elimination (RFE) in conjunction with Ridge Regression, we iteratively select a subset of features that exhibit the strongest predictive power, thereby enhancing model interpretability and performance.

The constructed predictive model undergoes rigorous evaluation, with metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score serving as benchmarks for assessing its efficacy. By comparing the performance of different model iterations and feature combinations, we aim to ascertain the optimal configuration that best captures the complex dynamics of the housing market.

Furthermore, this study explores the impact of additional derived attributes, such as the Property Tax Burden, on housing prices. By integrating domain knowledge and external sources, we enrich the feature space, potentially enhancing the predictive capabilities of the model.

In summary, this paper endeavors to contribute to the field of housing market analysis by proposing a robust predictive modeling framework. Through an empirical examination of diverse features and methodologies, we strive to provide insights into the underlying determinants of housing prices, thereby empowering stakeholders with valuable information for decision-making and planning purposes.

RESEARCH METHODOLOGY

This study employs a multifaceted approach combining feature selection techniques with regression analysis to investigate housing price prediction. Initially, exploratory data analysis techniques such as histogram and box plot visualization are utilized to understand the dataset's characteristics and relationships. Subsequently, Recursive Feature Elimination (RFE) in tandem with Ridge Regression is employed to select the most relevant features and mitigate multicollinearity issues. Additionally, correlation analysis is conducted to identify potential attributes influencing housing prices. The methodology also involves the derivation of a new attribute, the Property Tax Burden, and its correlation with the target variable. Finally, linear regression models are trained and evaluated using various feature subsets to assess predictive performance. This comprehensive methodology aims to provide insights into the factors driving housing prices while offering robust predictive models for practical applications.

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 Data and Sources of Data

The data used in this study were sourced from the Boston Housing Dataset, a widely used dataset in machine learning and statistics. This dataset contains information collected by the U.S. Census Service concerning housing in the area of Boston, Massachusetts. The dataset was originally published in 1978 by Harrison, D., and Rubinfeld, D.L.

The primary source of the data is the U.S. Census Service, which collects comprehensive information on various aspects of housing, including crime rates, zoning proportions, property tax rates, and median housing prices. The dataset comprises 506 observations, each representing a different housing unit in the Boston area.

The dataset has been widely used in research and academia to explore various aspects of housing economics, urban planning, and predictive modeling. It provides a rich source of information for understanding the factors influencing housing prices and patterns in urban areas. In this study, the dataset serves as the foundation for analyzing the relationship between different housing attributes and median housing prices, as well as for developing predictive models to estimate housing prices based on these attributes.

1.3 Theoretical framework

The theoretical framework of paper lays out the foundational concepts and principles guiding our research. In the context of predicting housing prices using machine learning, it encompasses economic theories shaping housing markets, regression analysis fundamentals, feature engineering techniques, regularization methods like Ridge Regression, and the broader implications of data-driven decision-making in real estate. This framework provides the conceptual backbone for understanding the methodologies and insights driving your study.

Equations

1.Linear Regression Equation:

$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\varepsilon$

where y is the predicted housing price, $\beta0,\beta1,...,\beta n$ are the coefficients of the features $x1,x2,...,xn$, and ε represents the error term..

2. Ridge Regression Objective Function:

$\text{Objective}=\sum i=1N(yi-y\hat{}i)2+\alpha\sum j=1p\beta j2$

where N is the number of samples, yi is the actual housing price for the i-th sample, y^i is tthe predicted price, p is the number of features, βj are the coefficients, and α is the regularization parameter.

3. Feature Engineering Equations:

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

- Deriving a new feature like the property tax burden: Property Tax Burden = TAX/ MEDV

- Normalizing features using standardization or scaling: $x_{scaled} = x-\mu/ \sigma$

II. FEATURE SELECTION AND MODEL DEVELOPMENT

2.1 Objective:

"The goal of RFE is to iteratively select a subset of features that contribute the most to the model's performance."
- This is clear and accurate.

2.2 Procedure:

"RFE starts with the full set of features and fits the model." - Consider specifying the model type (Ridge Regression) here for clarity.

"It ranks the features based on their importance or coefficients." - This could be expanded slightly to explain that importance is typically measured by coefficients in the context of Ridge Regression.

"The least important feature(s) are removed, and the model is refitted with the remaining features." - This is accurate, but you could mention that the number of features removed in each iteration is determined by the step parameter in RFE.

2.3 Selection Criteria:

"The selection of features is typically based on their importance in the model, which can be measured by coefficients in the case of linear models like Ridge Regression." - This is clear and accurate.

2.4 Benefits:

"RFE helps in dimensionality reduction, selecting only the most relevant features for the model." - Clear and accurate.
"Ridge Regression is particularly useful in scenarios with multicollinearity, where RFE can assist in identifying a subset of features that contribute most to the prediction while mitigating collinearity issues." - This is clear, but you could briefly mention why Ridge Regression specifically helps with multicollinearity (because of the L2 regularization term).

Implementation:

"In scikit-learn, we can implement RFE with Ridge Regression using the RFE class in conjunction with the Ridge model." - This is clear and accurate.

"The number of features to select is specified with the n_features_to_select parameter." - Correct, but you could mention that if this parameter is not specified, RFE selects half of the features by default.

III. Results and Discussion:

1. **Feature Selection using RFE with Ridge Regression**:

- **Selected Features:**

The RFE with Ridge Regression selected 10 features: 'CRIM', 'CHAS', 'NOX', 'RM', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT'.

- **Mean Squared Error (MSE):**

The MSE achieved with the selected features using RFE with Ridge Regression was approximately 25.42.

2. **Linear Regression Model Evaluation for Different Feature Sets:**

- **Feature Sets Evaluated:**

Four different feature sets were evaluated: df_2, df_3, df_4, and df_5, along with the original feature set.

- **Evaluation Metrics:**

For each feature set, Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score were calculated.

Feature Set	MAE	MSE	R2 Score
df_2	3.333138247718127	27.160410178015706	0.6296330421323364
df_3	3.353010162999416	27.296304930181293	0.6277799432424498
df_4	3.332606157267546	27.894519248576305	0.6196225253019736
df_5	3.353836437480258	27.79409540705963	0.6209919329227032

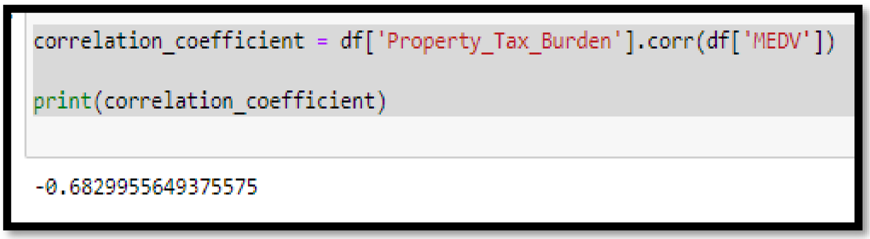
3. **New Attribute Creation and Analysis:**

- **New Attribute Created:**

A new attribute **Property_Tax_Burden** was derived by dividing the 'TAX' by 'MEDV'.

- **Correlation with Target:**

The correlation coefficient between Property_Tax_Burden and MEDV was calculated to be value.



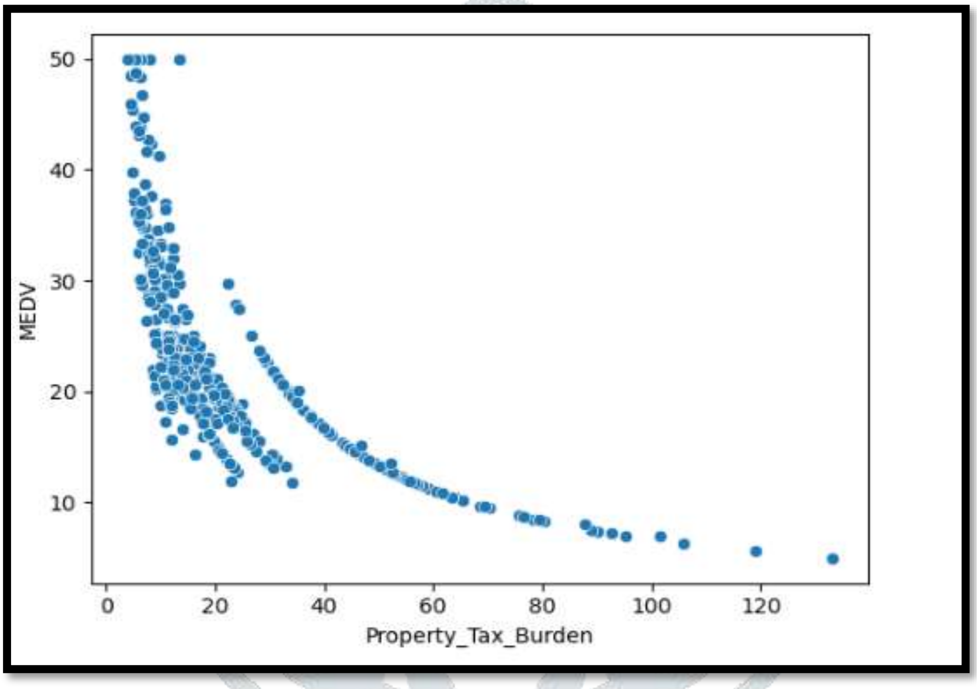
```
correlation_coefficient = df['Property_Tax_Burden'].corr(df['MEDV'])

print(correlation_coefficient)
```

```
-0.6829955649375575
```

- **Visualization:**

A scatter plot was created to visualize the relationship between **Property_Tax_Burden and MEDV**



4. **Linear Regression Model Evaluation with New Feature Set:**

- **Feature Set:**

A new feature set **df_6** was created including 'LSTAT', 'RM', 'PTRATIO', 'INDUS', and 'Property_Tax_Burden'.

- Evaluation Metrics:

MAE, MSE, and R2 Score were calculated for the linear regression model trained with df_6.

Feature Set	MAE	MSE	R2 Score
df_6	3.400083904985057	26.460874779450513	**0.6391721026910405**

5. Discussion:

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

5.2 Model Evaluation:
Evaluating multiple feature sets provided insights into which combination of features yields better predictive performance
.

5.3 New Attribute:
The creation of a new attribute allowed for the exploration of additional factors potentially influencing the target variable.

5.4 Model Performance:
Comparing model performance with different feature sets and the inclusion of a new attribute helps in understanding the impact of different variables on the model's predictive power.

5.5 Future Steps:

- Further refinement of feature selection and exploration of new attributes could lead to improved model performance.
- Additionally, considering more advanced modeling techniques or ensemble methods might enhance predictive accuracy further.