Predictive Modeling for Home Renovations Presentation

by Jacinta Mukii

INTRODUCTION

PROJECT OVERVIEW

For this project, I have used linear regression modeling to analyze house sales in a northwestern county and provide insights on house prices so as to help stakeholders understand the market dynamics and factors affecting housing values.

BUSINESS PROBLEM

Predicting house prices based on various housing features in the kc housing dataset. The task is to develop a predictive model using the provided dataset to provide homeowners with valuable advice on how home renovations can potentially increase the estimated value of their properties and by what magnitude.

MAIN OBJECTIVE

The main objective is to offer homeowners expert guidance on leveraging home renovations to enhance the estimated value of their properties. This objective seeks to empower homeowners with valuable insights regarding the potential increase in value resulting from specific renovation projects. By providing comprehensive advice, the goal is to enable homeowners to make well-informed decisions about their home renovations, effectively allocate their resources, and potentially maximize their return on investment when selling their homes.

Specific Objectives

01 Develop a simple linear regression model to **assess the relationship between the grade/reviews** of renovated houses and their prices, determining whether higher grades/reviews are associated with higher prices.

02 Evaluate the performance of a comprehensive regression model that **analyzes the relationship between all available renovation features** in the housing dataset and housing prices, with the goal of accurately predicting prices based on these features.

03 Build a multiple linear regression model to **investigate the relationship between the top 5 renovation features** in the housing dataset and housing prices, aiming to understand how these features collectively influence the prices.

Steps in model building and evaluation

Data Collection

This involves loading and inspecting the King County House Sales dataset in northwestern county.The dataset contains both categorical and numerical data.

Data Preprocessing

In this section, I have done the following : Dropping columns from the dataset, checking and removing duplicates in the dataset, detecting and dealing with null values in the dataset, separating values from words in a column.

Exploratory Data Analysis

I have summarized and analyzed datasets to gain insights, identify patterns and understand the underlying structire of the data using univariate, bivariate analysis and multivariate analysis.

Feature selection

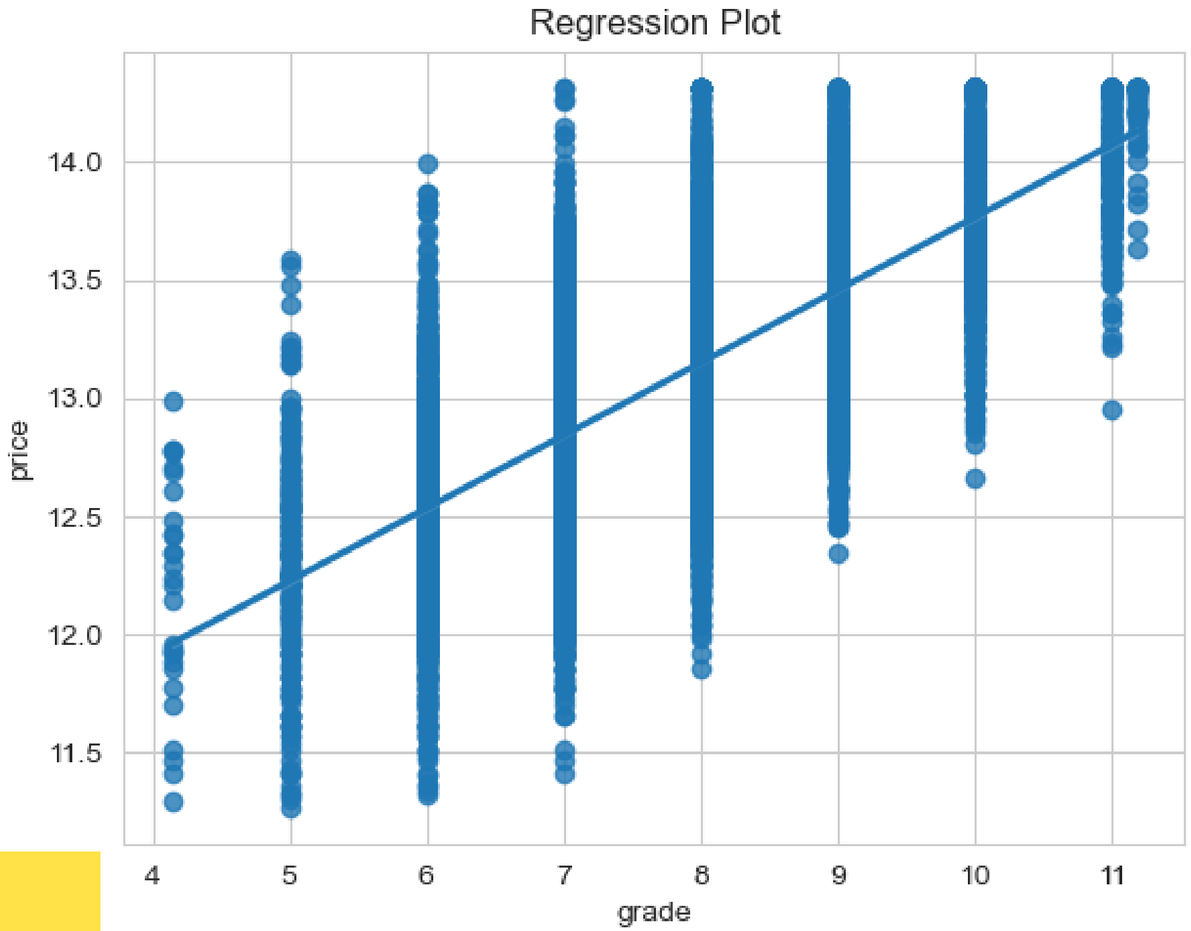
I performed multicollinearity between variables and splitting of data.

Data modelling and model performance evaluation

Modelling and evaluating the performance of linear regressions models.

Model interpretation

I have checked the best performing model using two metrics: r squared and mean squared error(mse).



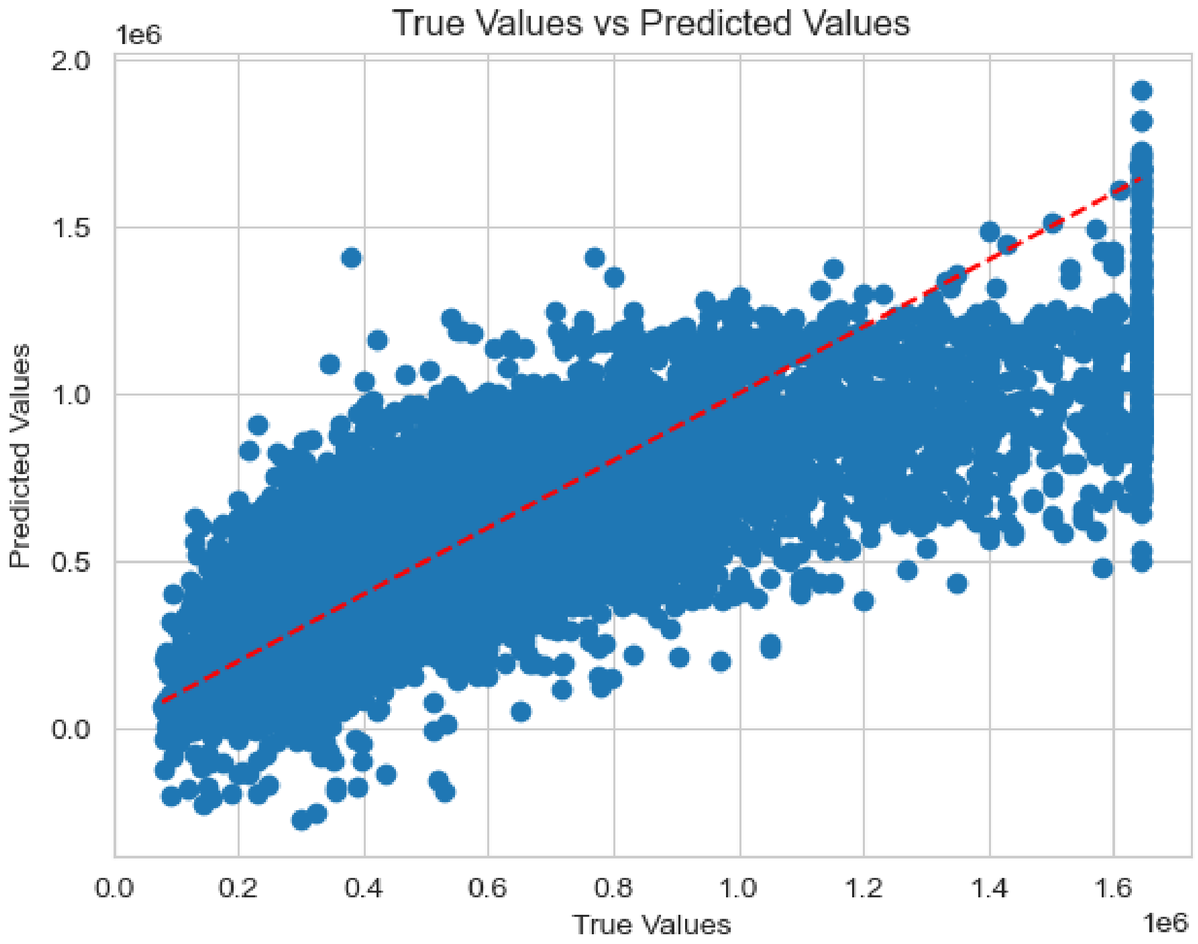
SIMPLE LINEAR REGRESSION MODEL WITH GRADE COLUMN AND PRICE

Linear regression for this model consists of price as the target variable and grade as the independent variable in the cleaned dataset.

The r squared for this model is 0.4941, approximately 49%.

This indicates that the model can effectively predict the price based on the grade/reviews by about 49%.

It suggests that 49% of the variability in the price can be explained by grade/reviews in the regression model.



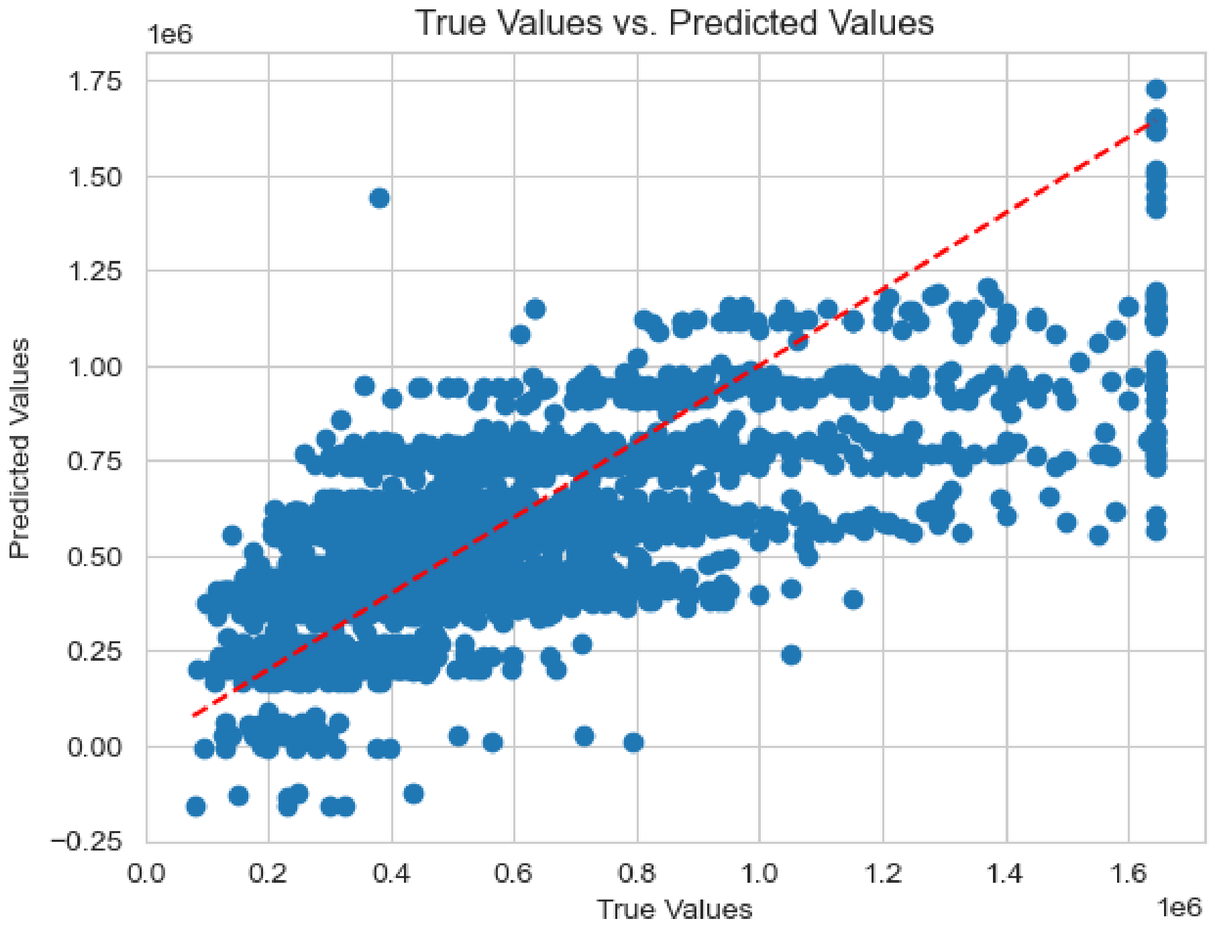
MODEL FOR ALL FEATURES AS THE INDEPENDENT VARIABLES

Linear regression for this model consists of all the feature variables in the cleaned dataset.

The r squared for this model is 0.6077, approximately 61%.

This indicates that the model can effectively predict the dependent variable based on the independent variables by about 61%.

It suggests that 61% of the variability in the price can be explained by all the feature variables/ independent variables in the regression model.



MODEL FOR THE TOP FIVE FEATURES IN THE DATASET

Linear regression for this model consists of the top 5 feature variables in the dataset.

The r squared for this model is 0.4983, approximately 50%.

This indicates that the model can effectively predict the dependent variable based on the independent variables by about 50%.

It suggests that 50% of the variability in the price can be explained by the top 5 feature variables in the regression model.

INTERPRETATION OF THE MODELS

Upon evaluating linear regression models, it is essential to compare their performance and select the final model based on the most favorable evaluation metrics.

I have used to 2 metrics: r squared and mean squared error for the models.

R squared -A higher R-squared value is generally preferred for a good model as it represents the proportion of the dependent variable's variance that is explained by the independent variables in the regression model, that is, it indicates that the model explains all the variability of the dependent variable.

Mean squared error - It measures the average squared difference between the predicted and actual values of the dependent variable. A lower MSE indicates that the model has smaller prediction errors and is better at fitting the data.

MODEL SELECTION

After careful evaluation, I have selected the 2nd model as the preferred choice.

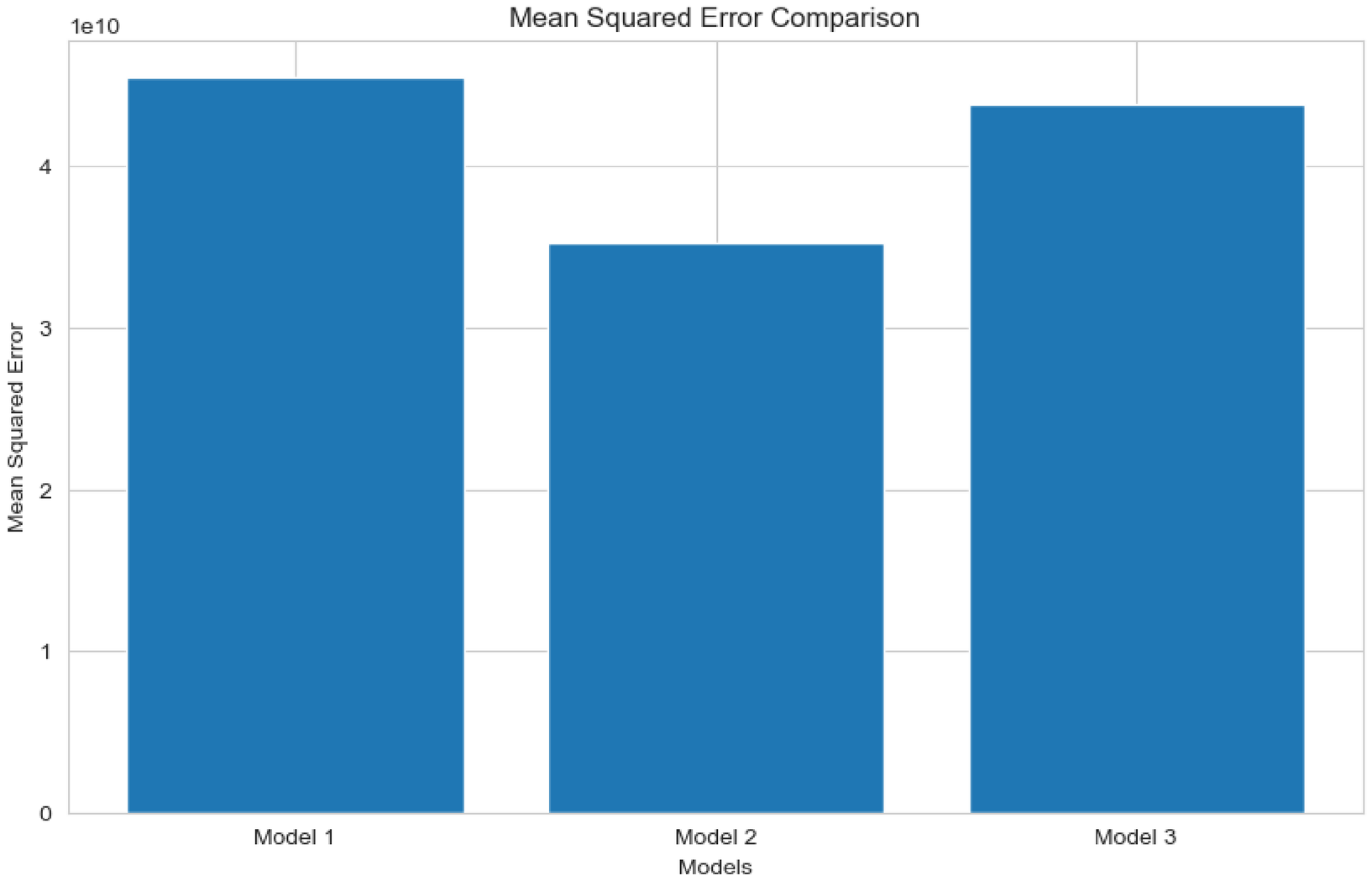
This decision is based on several factors, including its lower MSE (Mean Squared Error) and higher R-squared value compared to all other models.

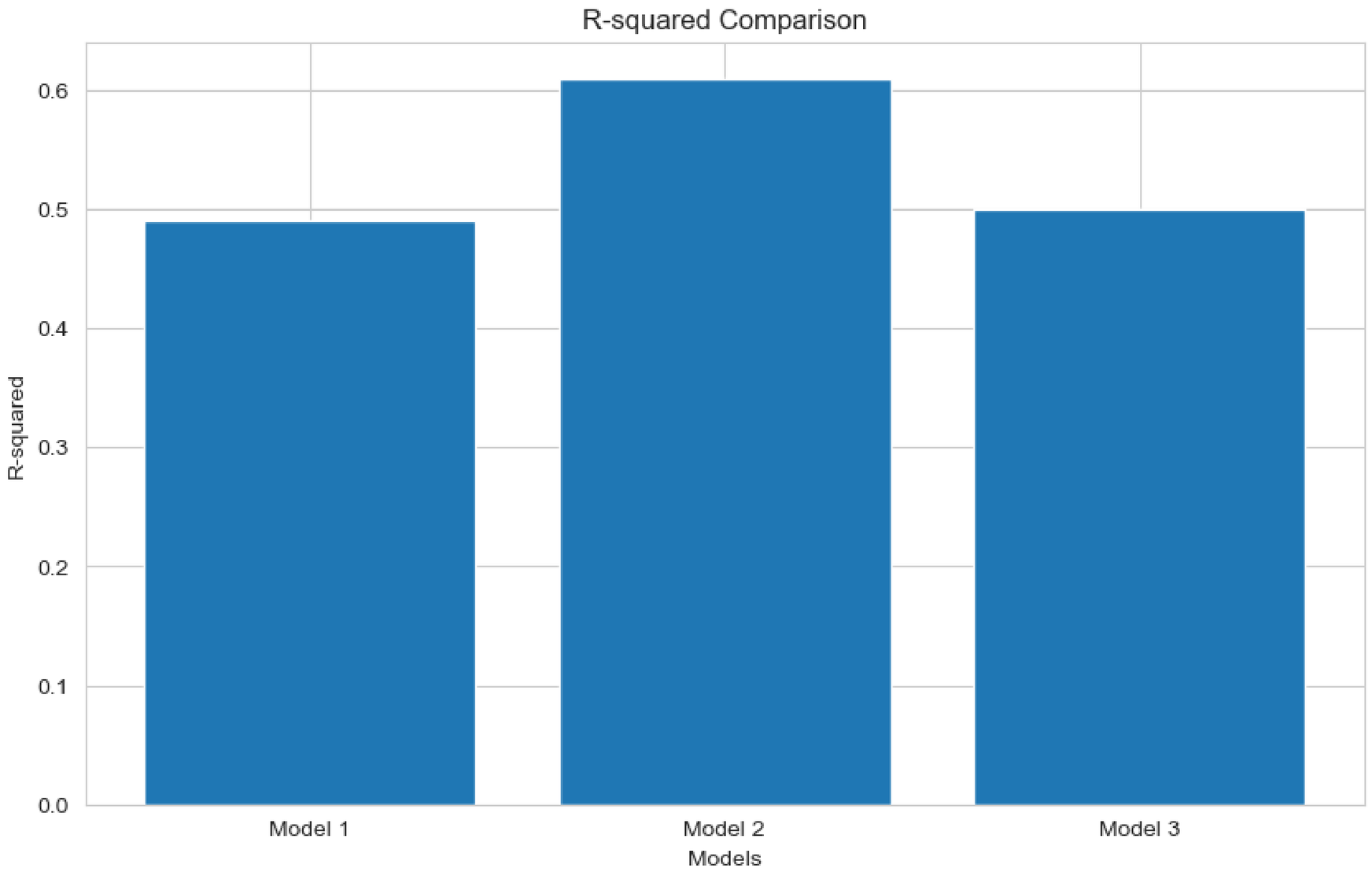
The lower MSE indicates that the 2nd model's predictions are closer to the actual values, reflecting its improved accuracy. Additionally, the higher R-squared value indicates that a larger proportion of the variance in the dependent variable is explained by the 2nd model.

Overall, these performance metrics suggest that the 2nd model provides a better fit to the data and is more capable of capturing the underlying patterns and relationships.

The model's performance is suboptimal, as indicated by an R-squared value of 0.6077. This means that only 61% of the variation in price is explained by the model.

Below is a graphical representation.


Mean Squared Error Comparison
1e10
4
3
2
1
0
Mean Squared Error
Model 1
Model 2
Model 3
Models


R-squared Comparison
R-squared
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Model 1
Model 2
Model 3
Models

CONCLUSION

Stakeholders should leverage multiple housing renovation features to positively impact housing prices.

By incorporating a variety of desirable renovation elements:

- such as high-quality materials,
- modern design elements,
- energy-efficient upgrades,
- and functional layouts,

Stakeholders can enhance the perceived value of properties and potentially command higher housing prices.

It is important to consider a comprehensive approach to renovations, taking into account the preferences and demands of potential buyers or renters, market trends, and the overall quality and attractiveness of the renovated properties.

RECOMMENDATIONS

1. I would recommend that stakeholders consider incorporating many housing features that have a significant impact on renovations.

- These features may include renovation grade/reviews, renovation floors, and other relevant factors. By considering these aspects, stakeholders can gain a comprehensive understanding of the renovation landscape and make informed decisions about the most effective strategies for enhancing the value of their properties.

2. Conducting surveys to capture clients' preferences regarding renovations related to the business problem can provide valuable insights.

- By actively seeking feedback from potential buyers or renters, stakeholders can better understand their specific renovation preferences, priorities, and expectations.

Contact details

Linked in profile :

Jacinta Mukii

EMAIL ADDRESS

jacinta.mukii@student.moringaschool.com

THANK YOU