Introduction

Title: Insightful Predictions: Multiple Linear Regression Approach to House Price Modelling in R

Objective: To understand the factors affecting house prices using statistical analysis and predictive modeling techniques with the Ames Housing dataset.

Table of Contents

1. Introduction and Background
   - Overview and Problem Statement
   - Literature Review
2. Methodology
   - Analytical Approach and Tasks
   - Data Exploration and Data Quality Assessment
   - Variable Selection
   - Data Quality Issues
   - Addressing Data Quality Issues
   - Hypothesis Testing
   - Regression Model Techniques
   - Model Building
3. Results and Discussion
   - Presentation of Key Outputs
   - Presentation of Key Outputs of All Models
   - Plot of Key Outputs of Model 3
   - Model Assumptions
4. Reflective Commentary
   - Further Steps
   - Learnings and Future Aspiration

5. References
6. Appendix

1. Introduction and Background

1.1 Overview and Problem Statement

This task focuses on investigating house pricing dynamics using statistical analysis and predictive modeling techniques with the Ames Housing dataset. The main objective is to understand the factors that influence house prices by examining various house attributes.

1.2 Literature Review

The literature review involved analyzing approximately 10 research publications focused on using machine learning algorithms for predicting residential property values. The table in the document summarizes these papers, highlighting the models used and their accuracy.

2. Methodology

2.1 Analytical Approach and Tasks

The analytical process includes data pre-processing, hypothesis formulation, data visualization, statistical association measurement, regression analysis, and model evaluation. The approach follows the conventional methodology employed in Machine Learning models, similar to CRISP-DM.

2.2 Data Exploration and Data Quality Assessment

Data exploration involves analyzing and characterizing the given data to assess its quality. The dataset comprises 78 variables, with the target variable being sale_price. Built-in functions in R such as summary(), count(), and is.na() were used for this purpose.

2.3 Variable Selection

A total of 17 variables were selected based on three factors: hypothesis, research papers, and logical reasoning. These variables include lot_area, neighbourhood, frontage, year_remod, room_tot, zone, year_built, half_bath, full_bath, bedrooms, aircon, kitchen, foundations, stories, heat_type, house_quality, and house_condition.

2.4 Data Quality Issues

The identified data quality issues include outliers and missing values in variables like lot_area, frontage, year_built, and house_quality. The document provides a detailed table of these issues.

2.5 Addressing Data Quality Issues

Data quality issues were addressed using techniques such as the Inter-Quartile Range (IQR) to remove outliers and the filter function from the “dplyr” library in R.

2.6 Hypothesis Testing

Hypothesis testing involved evaluating the null hypothesis (H0) against the alternative hypothesis (Ha) for various variables like lot_area, neighbourhood, room_tot, year_remod, and frontage. The statistical analysis provided

significant evidence to support the relationships between these variables and sale_price.

2.7 Regression Model Techniques

Multiple linear regression was used to build the models. The analysis aimed to ascertain the relationships between multiple variables and the dependent variable (sale_price).

2.8 Model Building

The forward approach was used to construct four models:

- Model 1: Based on variables derived from the hypothesis.
- Model 2: Based on variables derived from the hypothesis and literature review.
- Model 3: Including all variables (hypothesis, literature review, logical reasoning).

3. Results and Discussion

3.1 Presentation of Key Outputs

- **Model 1**: Adjusted R-squared of 0.7236, indicating 72.36% variance explanation in sale_price.
- **Model 2**: Better fit with an adjusted R-squared of 0.7627.
- **Model 3**: Highest adjusted R-squared of 0.8651 and lowest residual standard error, indicating better accuracy.

3.2 Presentation of Key Outputs of All Models

The table in the document presents the Root Mean Squared Error (RMSE), R-squared, and Mean Absolute Error (MAE) for all three models, showing Model 3 as the most accurate.

3.3 Plot of Key Outputs of Model 3

Graphs for Model 3 outputs are provided to visually demonstrate the relationships and accuracy.

3.4 Model Assumptions

The assumptions of independence, multicollinearity, and residuals are discussed. The Durbin-Watson test indicates no severe autocorrelation, and the Variance Inflation Factor (VIF) suggests no severe multicollinearity.

4. Reflective Commentary

4.1 Further Steps

Future steps include deploying the models across different organizational segments using R-Shiny to create user interfaces, providing actionable insights for strategic decision-making.

4.2 Learnings and Future Aspiration

This module enhanced proficiency in libraries like CARET, LM, TIDYVERSE, and GGPLOT for creating complex linear regression models. The aspiration is to contribute to machine learning, particularly in supervised learning algorithms,

and to use advanced ML algorithms to improve predictive modeling and business strategies.

5. References

A detailed list of references is provided, including sources from statistical analysis, machine learning, and housing price prediction literature.