Summary Report of the Multiple Regression Model

Introduction

This report outlines the development and evaluation of a multiple regression model designed to predict house prices based on various features, including size, number of bedrooms, age, and proximity to downtown. The dataset used comprises multiple observations of house characteristics and their prices.

Data Preparation

The dataset was examined for missing values, which were filled with column means. Numerical features were standardized for consistent scaling, essential for regression model performance. There were no categorical variables in this dataset that required one-hot encoding.

Model Training and Evaluation

The dataset was divided into training (70%) and testing (30%) sets. The LinearRegression model from Scikit-learn was employed.

Functions Used

- train_test_split: Randomly splits the dataset into training and testing sets, with a specified test size.
- LinearRegression: Implements ordinary least squares linear regression, fitting a linear model that minimizes the residual sum of squares.
- fit: Trains the model on the training data, adjusting coefficients to reduce prediction errors.
- predict: Generates predictions based on the test dataset features.

Evaluation Functions

- mean_squared_error: Computes the mean squared error between actual and predicted values, reflecting average squared differences.
- r2_score: Calculates the coefficient of determination, indicating the proportion of variance in the dependent variable explained by independent variables.
- Adjusted R-squared: Adjusts the R-squared value based on the number of predictors, providing a more accurate goodness of fit measure.

Performance Metrics

The model's performance was assessed using:
- Mean Squared Error (MSE): A lower MSE reflects better performance, indicating smaller prediction errors.
- R-squared: Ranges from 0 to 1, showing the proportion of variance explained by the independent variables; a higher value indicates a better fit.
- Adjusted R-squared: This metric accounts for the number of predictors, penalizing irrelevant additions, thus offering a reliable performance measure.

Results showed an MSE of 0.00, an R-squared of 1.00, and an Adjusted R-squared of 1.00, suggesting the model effectively predicts house prices based on the selected features.

Interpretation of Model Coefficients

The regression coefficients indicate the expected change in house price for a one-unit increase in each feature, holding others constant. Examples include:

- Size (sqft): Larger houses are associated with higher prices.
- Bedrooms: More bedrooms correlate with higher prices.
- Age: Older houses generally have lower prices, reflecting depreciation.
- Proximity to Downtown: Greater distances from downtown correlate with lower house prices.

These coefficients illustrate the significance of each predictor in the model.

Visualization of Predicted vs. Actual Prices

A scatter plot visualized the relationship between predicted and actual house prices, with points clustered near the diagonal line (y=x) indicating accurate predictions. Deviations from this line illustrate prediction errors.

Conclusion

The multiple regression model demonstrates a reasonable ability to predict house prices based on the selected features. Evaluation metrics indicate a satisfactory fit, and coefficient interpretation highlights each predictor's significance. Future work could explore additional features, regularization techniques, or more complex models to enhance predictive accuracy.