REGRESSION ANALYSIS : CIA REPORT

PAPER: REGRESSION MODELLING

NAME : NILANJANA DEY

REGISTER NO: 23122024

CLASS : MSC-DS-A

TOPIC : MULTIPLE LINEAR REGRESSION

INTRODUCTION

THEORITICAL BACKGROUND :

Multiple Linear Regression (MLR) is a statistical technique that models the relationship between a dependent variable and two or more independent variables by fitting a linear equation to the observed data. The general purpose of MLR is to understand the relationship between the dependent variable and the independent variables and to predict the dependent variable based on new observations of the independent variables.

Dependent Variable (Y): The variable you are trying to predict or explain.

Independent Variables (X1, X2, ..., Xn): The variables you are using to predict the dependent variable.

Regression Coefficients (β0, β1, ..., βn): The parameters of the model that represent the relationship between each independent variable and the dependent variable.

The MLR Equation :

The MLR model is represented by the following equation:

$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\epsilon$

Where:

Y is the dependent variable.

X1,X2,…,Xn are the independent variables.

β0 is the intercept (constant term).

β1,β2,…,βn are the coefficients for each independent variable.

ϵ is the error term (residuals).

Assumptions of MLR :

Linearity: The relationship between the dependent variable and each independent variable is linear.

Independence: The residuals (errors) are independent.

Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.

Normality: The residuals are normally distributed.

No Multicollinearity: The independent variables are not highly correlated with each other.

Steps to Perform MLR :

Data Preparation: Gather and clean the data, ensuring no missing values or outliers.

Exploratory Data Analysis (EDA): Understand the relationships and distributions of the variables.

Split the Data: Divide the data into training and testing sets.

Fit the Model: Use the training data to fit the MLR model.

Evaluate the Model: Use the testing data to evaluate the model's performance using metrics such as RMSE, R^2, adjusted R^2, etc.

Residual Analysis: Check the assumptions of MLR by analysing the residuals.

DATA PREPARATION

Dataset : https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data

Data Description :

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Performing EDA :

Checking data type of each column

Checking for missing values

Checking for duplicates

Checking for correlation amongst the features

Choosing the target variable and the independent features

AIM / OBJECTIVE :

To fit a Multiple Linear Regression Model to predict the House Prices in King's County.

According to the Correlation Plot :

Independent variables : 'bathrooms', 'sqft_living', 'grade', 'sqft_above', 'sqft_living15'

The above independent variables are selected after checking the multicollinearity using VIF values. The VIF values for the above features is <5.

VIF (Variance Inflation Factor) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, leading to inflated standard errors and potentially misleading interpretations of the regression coefficients. VIF quantifies how much the variance of an estimated regression coefficient is increased because of multicollinearity.

Interpreting VIF Values:

VIF = 1: No multicollinearity. The variance of the coefficient is not inflated at all.

VIF > 1 and < 5: Moderate multicollinearity. The variance of the coefficient is moderately inflated.

VIF > 5: High multicollinearity. The variance of the coefficient is significantly inflated, indicating a problem with multicollinearity.

ANALYSIS

1) **Feature Scaling** :
   Feature scaling is a preprocessing step in machine learning that standardizes or normalizes the range of independent variables or features in the dataset. This is important because many machine learning algorithms perform better or converge faster when the features are on a similar scale. There are two common methods for feature scaling:

   Standardization (Z-score normalization):

   Standardization scales the features to have a mean of 0 and a standard deviation of 1.
   It preserves the shape of the original distribution and does not bound the values to a specific range.

It is less affected by outliers compared to min-max scaling.

Formula:

$z=(x-\mu)/\sigma$

, where

μ is the mean and

σ is the standard deviation.

2) **<u>TRAIN TEST SPLIT</u>**:
   The train-test split is a common technique used in machine learning to divide a dataset into two subsets: one for training the model and one for testing the model's performance.

3) **<u>FITTING THE MODEL</u>**
   We fit the linear regression line and calculate the intercept and regression coefficients.

4) We calculate the Predicted values based on the test data.
5) **<u>MODEL EVALUATION METRICES</u>**

   a) RMSE (Root Mean Squared Error) is a commonly used metric to evaluate the performance of a regression model. It measures the average magnitude of the errors between predicted and actual values. Lower RMSE values indicate better model performance.

      RMSE: 0.034298761112248326

   b) MAE (Mean Absolute Error) is another commonly used metric to evaluate the performance of a regression model. It measures the average absolute difference between the predicted and actual values. Like RMSE, lower MAE values indicate better model performance.
      MAE: 0.021590203635828937

   c) R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the

independent variables in a regression model. It ranges from 0 to 1, with higher values indicating better fit of the model to the data.
R-squared: 0.5481017682465567

d) Adjusted R-squared is a modified version of the R-squared value that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not improve the model's performance. Adjusted R-squared typically provides a more accurate measure of the model's goodness of fit when comparing models with different numbers of predictors.

   Adjusted R-squared: 0.5475783744177942

e) MAPE : MAPE (Mean Absolute Percentage Error) is a metric used to evaluate the accuracy of a forecasting model. It measures the average absolute percentage difference between the predicted and actual values, relative to the actual values. MAPE is expressed as a percentage, and lower values indicate better accuracy.

   MAPE: 43.374164007973505

Interpretation of evaluation metrics:

The RMSE and MAE values almost tend to 0 which states that the fitting is quite good.

R squared and Adjusted R squared values are somewhat moderate, around 0.54, that is, the model can predict the relationship between the variables by 54%.

The MAPE value shows that the error percentage is around 43% showing that the fitting is not so great.

6) We calculate residuals to perform the Residual Analysis :
   - The residual vs fitted values plot shows that the points scattered around the residual line is not quite random and shows a few outliers.

- The Predicted values are somewhat equal to the actual values indicating that the fitting is not so great.
- **CHECK FOR NORMALITY OF THE ERRORS**

  Q-Q plot :

  A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a particular distribution, often the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the data follows the specified distribution.

  Interpretation : The errors are somewhat normally distributed as shown by the Q-Q plot and the histogram. The mean is around 0. Here also, we can spot certain outliers.
- **CHECK FOR HOMOSCEDASTICITY**

  Homoscedasticity is an important assumption in linear regression and other statistical models, indicating that the variance of the errors (residuals) is constant across all levels of the independent variables. When this assumption is violated (heteroscedasticity), the estimates of the coefficients can be inefficient, and the standard errors can be biased, leading to invalid statistical inferences.

  The **Breusch-Pagan test** is used to detect heteroscedasticity in a regression model. It tests whether the variance of the residuals from a regression is dependent on the values of the independent variables. If the test indicates heteroscedasticity, it suggests that the assumption of constant variance (homoscedasticity) is violated.

  Interpretation : Interpretation : Since the p-value is 0 we can understand that the error variance is not constant.
- **CHECK FOR AUTOCORRELATION**

  The Durbin-Watson (DW) test is a statistical test used to detect the presence of autocorrelation (serial correlation) in the residuals of a regression analysis. Autocorrelation can violate the assumptions of ordinary least squares (OLS) regression and lead to inefficient estimates.

Interpretation : Durbin-Watson Statistic value around 2 indicates no evidence of autocorrelation. Hence we can say that the errors are independent.

CONCLUSION

1) The independant feature columns should be have been transformed to get a normal distribution. This should have reduced the error and maintained homoscedasticity.

2) We could have performed principal component analysis and worked with few principal components to get a more compact fitting line to predict the house prices.

3) The current regression line is not a good fit as the residual analysis states that the errors are heteroscedastic.

REFERENCES

- Fox, J. (2019). *Regression diagnostics: An introduction*. Sage publications.
- Welsch, R. E., & Kuh, E. (1977). *Linear regression diagnostics* (No. w0173). National Bureau of Economic Research.
- https://www.geeksforgeeks.org/ml-linear-regression/
- https://www.scaler.com/topics/data-science/residual-analysis/