Name: Shreeyash Shripad Dongarkar

PRN: 22510025

TY BTech CSE

Machine Learning Lab

Assignment 7: Multiple Linear Regression

Initial Observations:

* Initial R2 is 0.130 which means model only explains 13% variance
* Has a multicollinearity problem due to points having high VIF
* Also contains influential points

High VIF means a variable is redundant making interpretation difficult. Influential points can significantly shift regression coefficients, leading to misleading conclusions. Removing highly correlated features and influential point improves model generalization and avoids overfitting.

Reducing Multicollinearity:

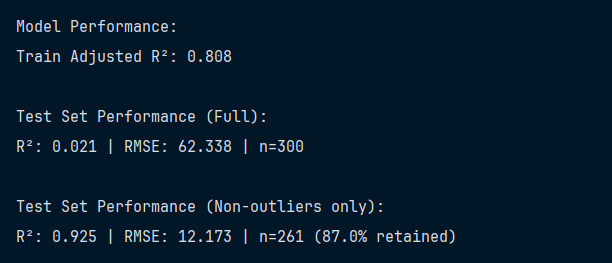
* To reduce multicollinearity, we iteratively calculated the Variance Inflation Factor (VIF) for each feature. In each iteration, the feature with the highest VIF value was removed, and the process was repeated until all remaining features had VIF values below a set threshold
* While doing this, we also considered the **p-values** of the features. A high VIF does not necessarily imply that a feature should be removed if it is **statistically significant**. Therefore, we retained variables with high VIF only if they had low p-values, indicating they were important predictors in the model.

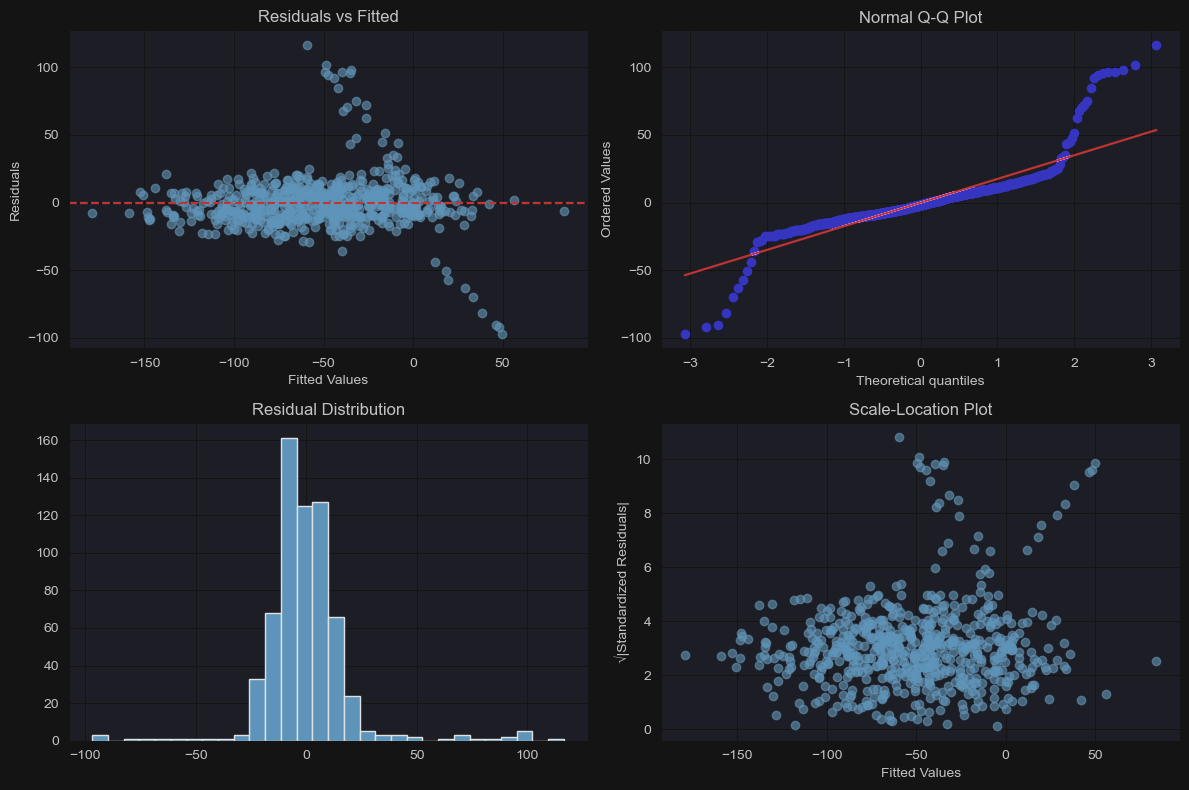
Hence, we removed features x4 and x11 as both had severe multicollinearity and high p-values meaning they are statistically insignificant. After this R2 was dropped slightly to 0.126.

Removal of Influential Points Using Cook’s Distance:

* First, we fit a temporary linear regression model and calculate residuals which measure how far each value is from actual value. To understand the variation in residuals we calculate MSE and a higher MSE implies model has more unexplained variance. Then we compute Leverage to measure how much influence a point has based on how far it is from mean of independent variables. h values closer to 1 mean that a point has high leverage and can significantly influence the model. Then we calculate Cook’s Distance to see if any point has high residual and high leverage. Then we remove them by keeping a threshold.

Found 56 outliers using this method which can distort regression coefficients, reduce generalizability and inflate error metrics.





Residuals are evenly spread around the 0 line but shows mild heteroscedasticity. From Q-Q Plot we can see a slight skewness at tails. Histogram is roughly symmetric implying approximate normality. Scale-Location Plot suggest slight increase in higher fitted values.