

Multicollinearity In Machine Learning: Detection, Impact, And Remedial Techniques

Dr. M. Chandrasekhar Reddy^{1*}, Dr. M. Naresh²

^{1*}Assistant Professor in Mathematics, Department of Basic Sciences & Humanities, B V Raju Institute of Technology, Narsapur, Medak, Telangana-502 313, India.

²Academic Consultant, Department of Statistics, Sri Venkateswara University, Tirupati 517502, Andhra Pradesh, India

Abstract

Machine learning and statistical models commonly experience multicollinearity effects because two or more independent variables within datasets shows strong interconnections. When this phenomenon occurs it weakens predictive models and their interpretability which results in untrustworthy coefficient estimates and larger standard error margins and decreased predictive capability. This paper investigates the factors that create multicollinearity together with detection methods and available solutions for this problem in the machine learning field. The effects on linear regression together with logistic regression and additional models such as decision trees and neural networks receive analysis. Researchers employ three approaches to identify multicollinearity that include Variance Inflation Factor (VIF) and correlation matrices as well as principal component analysis (PCA). We analyze multiple approaches for handling this problem which includes utilizing feature selection as well as implementing regularization methods such as Ridge and Lasso regression and performing dimensionality reduction.

The objectives of this research are three fold: (1) to analyze the significance of multicollinearity in statistical inference and machine learning models, with emphasis on its impact on error variances, feature selection, and overall predictive performance; (2) to evaluate and compare the effectiveness of detection methods for multicollinearity such as Variance Inflation Factor (VIF), correlation matrices, Principal Component Analysis (PCA), and eigenvalue decomposition across different modeling contexts; and (3) to validate through empirical analysis, using real-world datasets, how addressing multicollinearity improves model accuracy and interpretability, while also highlighting the trade-offs of methods like PCA in terms of interpretability.

The research employs benchmarked datasets through empirical assessment to show model performance impacts from multicollinearity alongside practical real-world solutions for its management. The research intends to establish thorough knowledge of multicollinearity while providing practitioners with practical methods to stabilize model accuracy and clarity.

Keywords: Multicollinearity, Machine Learning, Feature Selection, Regularization, Variance Inflation Factor, Principal Component Analysis, Model Interpretability.

Introduction

The statistical phenomenon of Multicollinearity demonstrates in datasets through related independent variables with high correlation that results in unstable model predictions while decreasing interpretability. The effect of multicollinearity becomes extremely problematic for regression-based models because it generates unstable coefficients that diminish statistical inference strength [1]. The work of Farrar and Glauber (1967) revealed that multicollinearity warps hypothesis testing to such an extent that it becomes hard to recognize independent predictor effects and yields incorrect evidence regarding variable significance [3]. The rise in multicollinearity becomes more probable when dealing with extensive datasets which leads to both over fitting with excellent training results but weak generalization to external data points [4]. Successful model development requires that multicollinearity be properly resolved.

The statistical method Principle Component Analysis (PCA) provides an effective solution to multicollinearity problems by converting related variables into independent

uncorrelated components which maintains crucial data elements without increasing dimensions [2]. PCA realizes its objective of data transformation through a process that project features onto orthogonal axes based on their respective explained variance order [5]. The data transformation under this approach manages multicollinearity issues then improves computational performance by selectively concentrating on key information in the data structure [6]. The accuracy of predictive models increases when using PCA on high-dimensional datasets because the technique removes useless data while decreasing signal noise [7]. The interpretable aspects of principal component results prove difficult due to their linear structure which combines original variables in unpredictable ways [8]. Although PCA has limited interpretability its combination with other methods including regularization makes it a preferred data preprocessing technique when dealing with multicollinearity [9].

Many researchers have investigated PCA's effectiveness in handling multicollinearity problems in finance and bioinformatics as well as social sciences fields [10]. Finance researchers apply PCA to decrease dimensional complexity of related financial signals which leads to more stable prediction models according to [11]. The application of PCA in bioinformatics becomes successful at detecting important molecular patterns by reducing multicollinearity problems when analyzing gene expression data [12]. PCA has multiple benefits yet underwent several important restrictions in its application. The algorithm depends on linear variable relationships but these assumptions might not apply to every dataset and it needs exact data scaling for proper result generation [13]. To achieve feature selection in addition to dimension reduction PCA requires the integration with Lasso regression techniques [14]. The research investigates how PCA treats multicollinearity while assessing its benefits alongside challenges and presenting step-by-step guidance for ML workflow application [15] [16].

2. Problem Statement

2. Why Multicollinearity is a Problem?

A regression model develops multicollinearity whenever two or more predictor variables demonstrate strong interdependency to the point it creates data redundancy. The predictive power of full models remains stable despite multicollinearity however this statistical condition deteriorates coefficient reliability thereby becoming an essential problem for machine learning systems and statistical models.\

2.1. Inflated Variance in Coefficient Estimates

High standard errors emerge when researchers analyze models affected by multicollinearity thus preventing proper determination of individual variable effects [17]. High levels of predictor variable correlation cause minor changes in the dataset to produce significant changes in estimated coefficients. Such conditions produce invalid conclusions and untrustworthy model explanations [18].

2.2. Reduced Interpretability of the Model

The main objective behind regression-based machine learning models is determining the impact of each predictor variable on the results. When multiple predictors have strong relationships it becomes impossible to determine their individual influences during analysis because their combined effects may become distorted [19]. The situation becomes challenging in fields like finance and healthcare and economics because decision-making depends significantly on interpretability [20].

2.3. Instability in Model Predictions

Data over fitting linked to highly correlated variables brings about failed generalization abilities that affect new data predictions [21]. The training data sensitivity of the model becomes excessive because it leads to erratic prediction outputs [22]. The stability of linear regression models becomes possible thanks to Ridge and Lasso regression which apply penalties for coefficient estimate shrinkage [23].

2.4. Difficulty in Statistical Inference

The variance of regression coefficients increases due to multicollinearity which produces high p-values that reduce statistical significance assessment capability [24]. The significance test demonstrates important variables becoming non-significant thus producing wrong model evaluation about predictor value [25]. The Variance Inflation Factor (VIF) provides a method for detecting and measuring multicollinearity through its analysis since values exceeding 10 usually signal severe multicollinearity [26].

2.5. Poor Feature Selection and Model Interpretation

Machine learning algorithms processing feature importance rankings such as decision trees and linear regression will frequently misrecognize the most vital predictors when multicollinearity exists [27]. Such conditions result in faulty feature selection practices through which useless variables enter analysis while vital variables get omitted [28]. The model performance improves through dimensionality reduction when features are transformed into orthogonal components through Principal Component Analysis (PCA) and Partial Least Squares (PLS) techniques [29].

2.6. Inefficiency in Model Training

When multicollinearity occurs within a dataset it adds unintentional redundancy that slows down the training process for machine learning models and makes them handle more unnecessary complexity [30]. The technique poses problems for big data collections that must maintain efficient operation [31]. The regularization methods Ridge and Lasso regression reduce correlated predictor effects and control large coefficient magnitudes [32].

3. Detection Methods

Several statistical detection methods exist to measure the degree of predictor variable correlation within a model structure. The following list presents popular statistics for multicollinearity diagnosis along with their respective references.

3.1. Correlation Matrix

The correlation matrix functions as an effective tool to detect multicollinearity by presenting all pairs of variable correlations [33]. High values of correlation exceeding 0.7 and 0.8 signal potential multicollinearity issues according to [34]. The detection of bivariate correlations stands as the main limitation because this technique fails to measure complex multicollinearity dynamics between multiple variables at once [35].

3.2 Variance Inflation Factor (VIF)

In practice we take R_i^2 as the proportion of the variance in the i^{th} independent variable that is correlated with the other independent variables in the model. It is an ideal instrument to test the collinearity between the i^{th} independent variable and the other modeled independent variables. The tolerance of the i^{th} independent variable i.e. $1 - R_i^2$ - proportion of variance it shares with the other independent variables in the analysis ($1 - R_i^2$) can now be defined. The form of the tolerance leads to a reduction in Variance Inflation Factor (VIF) which is the inverse of tolerance.

$$VIF = 1 / (1 - R_i^2)$$

The VIF contains huge conclusive statements in respect of the impact of R_i^2 on the variance of the estimated regression coefficient of the i th independent variable. The researchers will fail to apply rules or criteria appropriately on many occasions, which reveal when the values of VIF or tolerance have acquired largest values. Broadly speaking, we may state that the VIF of 10 or even 4 (which corresponds to the toleration level of 0.10 or 0.25) are indicators of too much or even serious multicollinearity. Although the model fitted is good, statistically, however, we can reject or define the model when VIF had the maximum values.

By the time the VIF attains such threshold levels, researchers are pursuing alternative approaches to decrease the multicollinearity and only few of them are explained below.

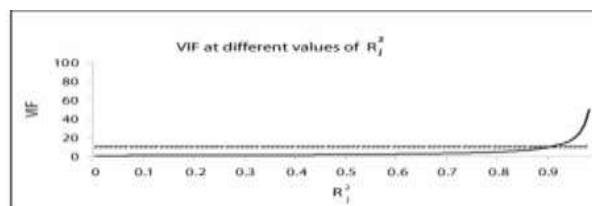
- i. Dropping any of the independent variables in their analysis that is displaying a high level of correlation with each other.
- ii. Using two or more independent variables to construct one index in such a way that we are able to eliminate the multicollinearity effects on the model building.
- iii. Getting close to the ridge regression as a method of analyzing multiple regression data afflicted with the issue of multicollinearity. In the case of multicollinearity, the least squares estimates will be unbiased, but a high variance implies that such estimates can be considerably different. Ridge regression minimizes the standard errors by introducing some form of bias with the regression estimates.
- iv. We can completely rebuild the model with new set of independent variables.

There are some risks or disadvantages of using some thumb rules about the VIF to decide the extent of multicollinearity issues in the data. Below we address some of these issues in detail.

- a. While checking the significance of regression coefficients, the conclusions based on VIF may mislead the results and the factor of stability is ignored completely.
- b. When VIF based conclusions are used in the model building, then there will be a small focus on the decision of rejecting or accepting the null hypothesis regarding the regression coefficients.
- c. Many researchers only study the remedies of multicollinearity and neglects the other causes of variation in the model.
- d. In practice, one would want to see the square root of the VIF rather than the VIF itself since the accuracy with which β_j is estimated is proportional to the standard error of β_j (not directly proportional to its variance). Due to its simplicity and straightforward interpretation, the VIF (or its square root) is the initial instrument of revealing the origins of imprecision.

No guidelines are given in determining the extent of variation inflation factors that bring about inefficient coefficients. Some authors recommend that largest VIF should be greater than value of 10, or that the mean VIF should be significantly larger than 1. Nevertheless, the latter values are quite arbitrary (Fox, 1992). $VIF = 10$ means that $R_j^2 = 0.9$.

Fig.3.2 Variance inflation factor



Variance Inflation Factor (VIF) is a measure of the extent, through which, multicollinearity inflates the variance of a regression coefficient [36].

A VIF greater than 5 or 10 indicates significant multicollinearity [37]. VIF is one of the most commonly used diagnostic tools in regression analysis [38].

3.3. Tolerance Index

Tolerance is the reciprocal of VIF and is defined as: $\text{Tolerance} = 1 - R^2$. A low tolerance value (below 0.1) suggests that a predictor variable is highly collinear with others, making it redundant in the model [39].

3.4 The Condition Index and Eigen Values

The second approach to test the level of multicollinearity is the size of the Eigen values of the correlation matrix of the regressors. The great variation between the eigenvalues means that there is a higher level of multicollinearity. Two characteristics of these Eigen values are of interest:

- Eigen values equals zero: there are very precise collinearities. Thus, the Eigen values that are very small show close linear relationships or high levels of multicollinearity.
- The square root of the largest to the smallest Eigen value.

$$K = \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{1/2},$$

is the more frequently used measure of the instability of the least-squares regression coefficients, and is sometimes called the condition number. A large condition number (e.g. 10 or more) implies that comparatively small variations in the data are likely to cause large variations in the least-squares estimate. The correlation matrix of the regressors is said to be ill conditioned in this event (Greene, 1993). The following is a simple situation but with two regressors model: the condition number is.

$$K = \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{1/2} = \left(\frac{1 + \sqrt{r_{12}^2}}{1 - \sqrt{r_{12}^2}} \right)^{1/2}$$

Setting K equal to 10 corresponds to $r_{12}^2 = 0.9608$.

The condition index, derived from the eigenvalues of the independent variable matrix, helps assess multicollinearity severity [40]. Condition indices above 30 indicate severe multicollinearity [41].

3.5 Farrar and Glauber Test

This statistical test examines the presence of multicollinearity by analyzing the determinant of the correlation matrix and conducting chi-square tests [42]. However, it is not widely used due to its computational complexity in large datasets [43].

Farrar and Glauber (1967) proposed a way to diagnose the multicollinearity problem and they used three test statistics namely χ^2 , F and t for testing the multicollinearity in the given data. χ^2 - test is used for the detection of the existence and severity of multicollinearity in a function including several explanatory variables. F- test is used for the location of multicollinearity. t -test is used to detect the variables which are responsible for multicollinearity.

3.6. Klein's Rule

Klein's Rule suggests that if the individual correlation between two independent variables is high, but their combined correlation with the dependent variable is low, multicollinearity is present [44]. This method is mostly used in econometrics.

3.7. Principal Component Analysis (PCA)

PCA helps detect multicollinearity by transforming correlated variables into a set of uncorrelated principal components [45]. If the first few components explain most of the variance while others contribute negligibly, multicollinearity is likely present [46].

3.8. Ridge Trace Analysis

Ridge regression can be used to analyze the stability of regression coefficients as a regularization parameter (λ) increases. If coefficients fluctuate significantly, it suggests multicollinearity [47].

4. Data Analysis:

We can detect the multicollinearity using different machine learning algorithms. To explore these opportunities, we are using a subset of the dataset taken from the Housing Prices Competition on Kaggle. The House Prices dataset on Kaggle is one of the most popular and widely used datasets for practicing machine learning, especially regression modeling. It provides a rich and complex real-world scenario of predicting house prices based on multiple features, making it an excellent resource for both beginners and advanced data scientists to sharpen their skills. Below is a detailed overview of this dataset.

The dataset was originally compiled from real housing data in Ames, Iowa, and is a modern alternative to the well-known but limited Boston Housing Dataset. It contains 79 explanatory variables (features) that describe almost every aspect of residential homes and a target variable: Sale price. The training dataset has 1460 observations, while the test set has 1459 observations without the sale price (to be predicted).

The subset of the data that we will be using we can find in the following links as well.
<https://github.com/vikashkmd/Multicollinearity/tree/main>.

To analyze the data set, we used the python programming language. Python provides powerful libraries like Pandas, NumPy, and Scikit-learn, which make data cleaning, analysis, and modeling efficient. It has excellent data visualization tools such as Matplotlib and Seaborn to create clear and insightful graphs. It supports machine learning and deep learning frameworks like Tensor Flow and PyTorch, enabling advanced predictive analytics.

4.1 Detection of multicollinearity:

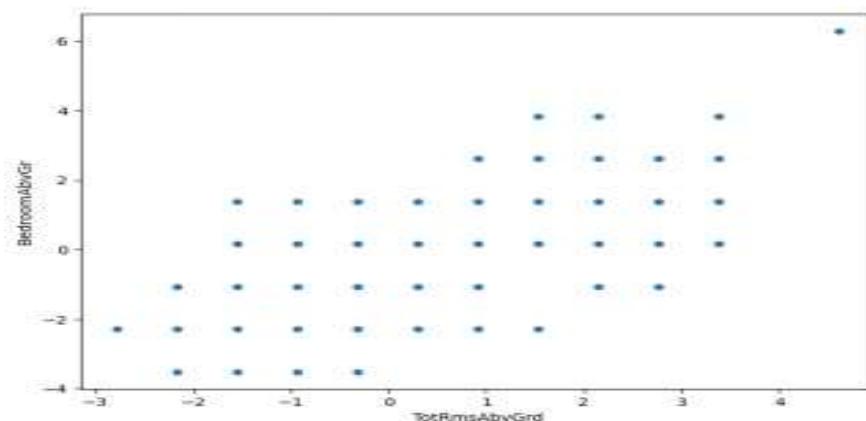
The first step in identifying the multicollinearity is to use the simple scatter plot matrix that will give broad idea about which variables are strongly correlated or not. The scatter plot matrix looks as follows in Fig: 4.1. Shows scatter plot matrix.

Fig 4.1 the scatter plot matrix



As we can see some positive correlations, we can explore them using two variable scatter plot in Fig. 4.2 shows two variables scatter plot.

Fig.4.2 two variables scatter plot



The above scatter plot show that “TotRmsAbvGrd” and “BedroomAbvGr” are having considerable positive correlation which might cause multicollinearity in the data. Fig.4.3 shows sales correlation heat map. The multicollinearity is also detected by using a correlation matrix.

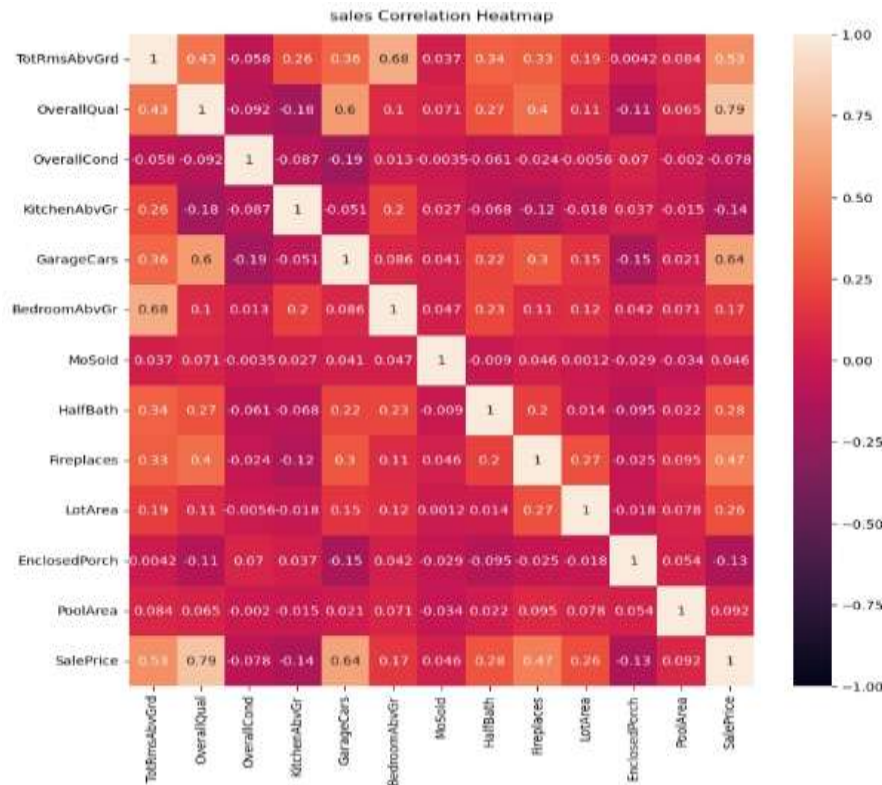
Fig. 4.3 Sales correlation heat map

Table 1: Shows variance inflation factor values. The measure of “Variance inflation factor” useful to find the multicollinearity in the data.

Table 1. Variance Inflation Factor

0	Tot Rms Abv Grd	51.1434
1	Over all Qual	32.3605
5	Bedroom Abv Gr	26.4810
3	Kitchen Abv Gr	20.3247
2	Over all Cond	17.6649
4	Garage Cars	11.1721
6	Mosold	6.3721
8	Fire Places	2.5530
9	Lot Area	2.3451
7	Half Bath	1.9053
10	Enclosed Porch	1.1758
11	Pool Area	1.0267

4.2 Treatment for Multicollinearity:

In order to have the multicollinearity in the regression models well under control, a number of techniques can be employed. These procedures assist in making sure that the model is correct and explainable, in the event that the independent variables are tightly connected.

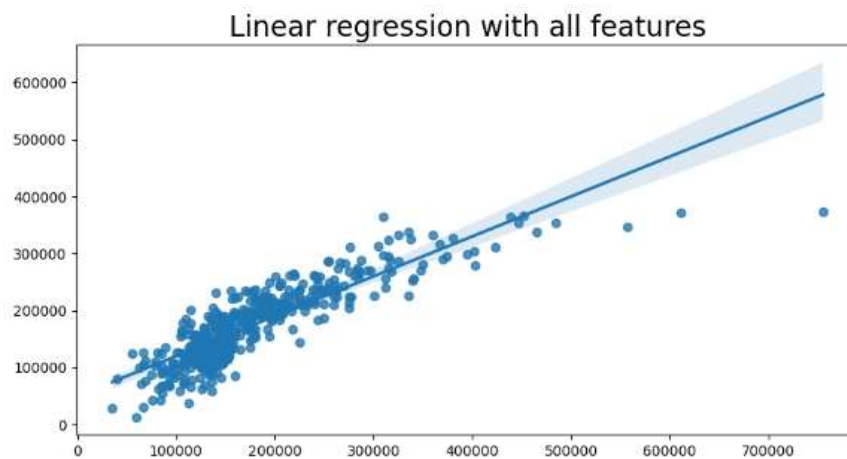
The easiest method of handling multicollinearity is to just drop one of the variables that are highly correlated, commonly that which has the largest VIF value. This is good but the disadvantage is that it may lead to loss of useful information unless done well. Performance metrics are indicated in Table 2.

Table 2. Performance metrics

RMSE	42489.1939
R^2	0.7413
Adjusted R^2	0.734
Cross Validation R^2	0.7166

The linear regression line also looks as follows

Fig.4.4 linear regression with all features

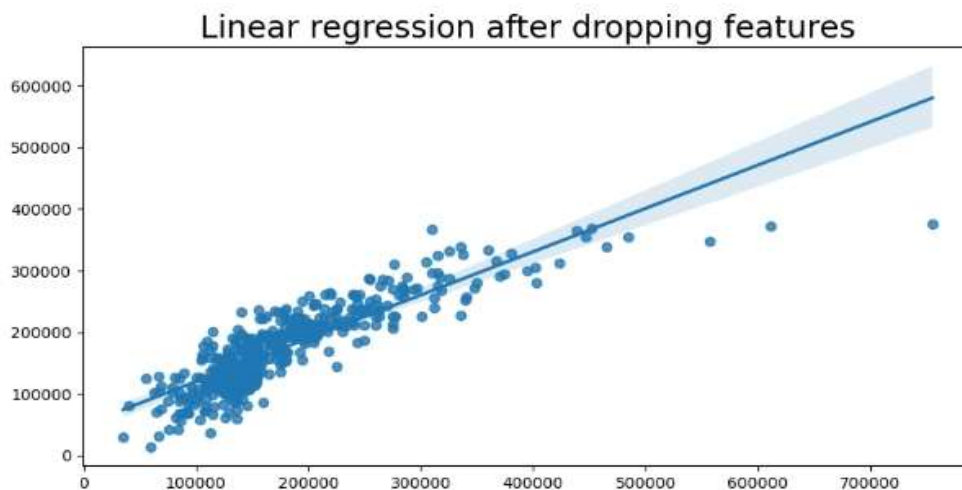


Now we will remove the two variables from the data.

`data1 = data.drop(['Overall Cond','MoSold','Pool Area'], axis=1)`

After that the regressions line looks as follows.

Fig.4.5 linear regression after dropping features



The performance measures are also slightly improved as follows.

Table 3. Improved performance measures shown

RMSE	42342.5517
R ²	0.7413
Adjusted R ²	0.7377
Cross Validation R ²	0.7311

This way multicollinearity can be addressed in the give data.

5. Conclusions And Future Recommendations:

5.1 Conclusions:

1. Multicollinearity is a key concept in statistical inference and machine learning, more specifically in multiple linear regression concept. Multicollienarity causes inflated error variances reduces model performances, and affects feature selection.
2. Detection methods such as Variance Inflation Factor (VIF), correlation matrices, principal component analysis (PCA), and eigenvalue decomposition provide effective ways to check for the existence of multicollinearity. Among these, VIF is widely used due to its simplicity, while PCA is effective in high-dimensional datasets.
3. In the presence of multicollinearity three are some remedial techniques, including regularization (Ridge and Lasso regression), feature selection, and dimensionality reduction, help mitigate the impact of multicollinearity. The Ridge regression stabilizes coefficient estimates, but we can see that Lasso regression eliminates irrelevant features.
4. From this current study, the empirical analysis using real-world datasets confirms that addressing multicollinearity improves model performance and interpretability. Removing highly correlated features enhances the predictive power of regression models.
5. Principle component analysis very effective in identifying the multicollinearity, but it has some limitations in terms of interpretability, as the transformed components do not directly correspond to the original variables.

5.2 Future scope for the study:

1. There is high scope to investigate how we can implement advanced ML algorithms like Deep Learning and can check on how multicollinearity is addressed with those models.
2. Developing hybrid feature selection methods that combine statistical techniques (e.g., VIF) with machine learning-based selection techniques to improve model performance.
3. We can extend our research to some of the non-linear models as well where multicollinearity is a serious issues there by applying techniques like tree-based models, ensemble learning, and deep neural networks.
4. We can automate the complete process of detection and treatment of multicollinearity with almost no human interventions is another future scope we can define in this context.
5. Conducting in-depth case studies across different domains like finance, bioinformatics, healthcare to understand how multicollinearity behaves in diverse datasets.

By addressing these areas, future research can enhance the robustness of machine learning models, improve decision-making, and create more interpretable AI systems.

References

1. James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013. An introduction to statistical learning. New York: Springer.
2. Jolliffe, I.T., 2002. Principal component analysis. 2nd ed. New York: Springer.
3. Farrar, D.E. and Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economics and Statistics*, 49(1), pp.92-107.
4. Hawkins, D.M., 2004. The problem of over fitting. *Journal of Chemical Information and Computer Sciences*, 44(1), pp.1-12.
5. Abdi, H. and Williams, L.J., 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp.433-459.
6. Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.
7. Zou, H., Hastie, T., and Tibshirani, R., 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), pp.265-286.
8. Bro, R. and Smilde, A.K., 2014. Principal component analysis. *Analytical Methods*, 6(9), pp.2812-2831.
9. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp.267-288.
10. Ringnér, M., 2008. What is principal component analysis? *Nature Biotechnology*, 26(3), pp.303-304.
11. Ledoit, O. and Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5), pp.603-621.
12. Alter, O., Brown, P.O., and Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18), pp.10101-10106.
13. Van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1), pp.1-15.
14. Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp.301-320.
15. Wold, S., Esbensen, K., and Geladi, P., 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), pp.37-52.
16. Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), pp.417-441.
17. Gujarati, D.N., Porter, D.C. and Gunasekar, S., 2022. Basic econometrics. 6th Ed. New York: McGraw-Hill.
18. Wooldridge, J.M., 2020. Introductory econometrics: A modern approach. 7th ed. Cengage Learning.
19. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. An introduction to statistical learning: with applications in R. 2nd ed. Springer.
20. Kutner, M.H., Nachtsheim, C.J. and Neter, J., 2020. Applied linear statistical models. 5th ed. McGraw-Hill.
21. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. Introduction to linear regression analysis. 6th ed. Wiley.
22. Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E., 2020. Multivariate data analysis. 8th ed. Pearson.
23. Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp.267-288.
24. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. and Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), pp.27-46.
25. Aiken, L.S., West, S.G. and Reno, R.R., 1991. Multiple regression: Testing and interpreting interactions. Sage.
26. O'Brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), pp.673-690.
27. Ghosh, S. and Vogt, A., 2012. Outlier detection and robustness in penalized regression. *Journal of Statistical Planning and Inference*, 142(5), pp.1428-1438.

28. Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp.301-320.
29. Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), p.20150202.
30. Fan, J. and Li, R., 2001. Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), pp.1348-1360.
31. Friedman, J., Hastie, T. and Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), pp.1-22.
32. Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12(1), pp.55-67.
33. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. *Introduction to linear regression analysis*. 6th ed. Wiley.
34. Wooldridge, J.M., 2020. *Introductory econometrics: A modern approach*. 7th ed. Cengage Learning.
35. Gujarati, D.N., Porter, D.C. and Gunasekar, S., 2022. *Basic econometrics*. 6th ed. McGraw-Hill.
36. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. *An introduction to statistical learning: with applications in R*. 2nd ed. Springer.
37. Kutner, M.H., Nachtsheim, C.J. and Neter, J., 2020. *Applied linear statistical models*. 5th ed. McGraw-Hill.
38. O'Brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), pp.673-690.
39. Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E., 2020. *Multivariate data analysis*. 8th ed. Pearson.
40. Belsley, D.A., Kuh, E. and Welsch, R.E., 2005. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.
41. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. and Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), pp.27-46.
42. Farrar, D.E. and Glauber, R.R., 1967. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1), pp.92-107.
43. Klein, L.R., 1955. *A textbook of econometrics*. Row, Peterson and Company.
44. Greene, W.H., 2018. *Econometric analysis*. 8th ed. Pearson.
45. Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), p.20150202.
46. Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp.301-320.
47. Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), pp.55-67.