International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Pearson's Correlation in Predictive Analytics and Machine Learning: Applications & Limitations

Jarita Das

Assistant Professor, Department of Statistics, Moinul Hoque Choudhury Memorial Science College, Hailakandi, Assam, India

Abstract

Pearson's correlation coefficient serves as a vital statistical measure in predictive analytics and machine learning by offering profound insights into linear relationships between variables. It is, thus, instrumental in understanding variable relationships, selecting features, detecting multicollinearity, and assessing the model performance. This paper explores the applications of Pearson's correlation in selecting features, reducing dimensionality, and interpreting the selected model. The paper highlights importance of Pearson's correlation in identifying suitable predictors and improving algorithmic performance in predictive analytics and machine learning. The paper also takes into account the limitations of Pearson's correlation that includes its sensitivity to outliers and reliance on assumptions of linearity and normality at the exclusion of non-linear associations. Alternative correlation measures like Spearman's rank and mutual information that address the shortcomings of Pearson's correlation are also taken within the purview of discussion. Through an examination of both the strengths and weaknesses of Pearson's correlation, the paper sheds light into use of Pearson's correlation in predictive modeling while stressing the need for adhering to complementary techniques in advanced machine learning applications.

Keywords: Pearson's correlation, predictive analytics, machine learning, Spearman's rank, Kendall's Tau

1. Introduction

Predictive analytics and machine learning, in the contemporary data driven world, have become indispensable to gain meaningful insights into diverse fields. The application of these tools can be widely seen in the fields like finance, healthcare, marketing, scientific investigations and environmental studies [1-3]. As concerns measures of linear dependence of variables, Pearson's correlation coefficient still remains one of the most widely used tools. Developed by Karl Pearson in the late 19th century, Pearson's correlation indicates the strength and direction of a linear association ranging from -1 to +1 [4]. The method is at once simple and suggestible, and hence extensively applied by statisticians and data scientists in exploratory data analysis to select features and validate models.

In predictive modeling, unearthing the associations between independent variables i.e. features and dependent variable i.e. the target variable is critical. Pearson's correlation assists data scientists to discern the features that are likely to be informative and thus aids in the process of feature selection. High correlation among input features may point out towards multicollinearity, an issue that seeks to undermine

the statistical significance of an independent variable [5]. Multicollinearity, can lead to distortion of model coefficients and reduce interpretability. The coefficient of Pearson's correlation, by signaling multicollinearity, plays a dual role; it highlights both useful predictors and potential issues within the dataset. Pearson's correlation, despite having wide utility, has inherent limitations. It effectively measures linear relationships but tend to overlook complex or non-linear interactions that are commonly present in real-world data. Moreover, Pearson's correlation coefficient is quite sensitive to outliers. Extreme values in a data set can unduly affect the value of Pearson's correlation coefficient and can potentially warp the strength and direction of the detected linear relationship. Outliers can, thus, can skew results and lead to deceptive conclusions. These limitations stipulate the need to take caution at the time of relying wholly on Pearson's correlation for analytical decisions.

In machine learning, where volume and complexity of data are likely to be high, the limitations set by Pearson's correlation become particularly significant. Models like neural networks, decision trees, support vector machines etc have the potential to capture patterns that linear statistics may fail to spot. Nonetheless, Pearson's correlation is a fundamental tool and can provide valuable preliminary insights. However, the assumptions and sensitivity of Pearson's correlation need to be managed carefully and it should be used alongside other tools and techniques for a comprehensive analysis.

2. Theoretical Background

Pearson's correlation coefficient (r) measures the degree of linear association between two variables, X and Y [6]. The formula is given by:

$$r = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y} \qquad (1)$$

where,

$Cov(X,Y)$ = Covariance between X and Y

$\sigma_X \sigma_Y$= Standard deviations of X and Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (2)$$

here, $\bar{x}\ and\ \bar{y}$ are the means of x and y respectively.

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^2\ and\ \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2$$

The range of Pearson's correlation coefficient (r) lies between -1 to +1,

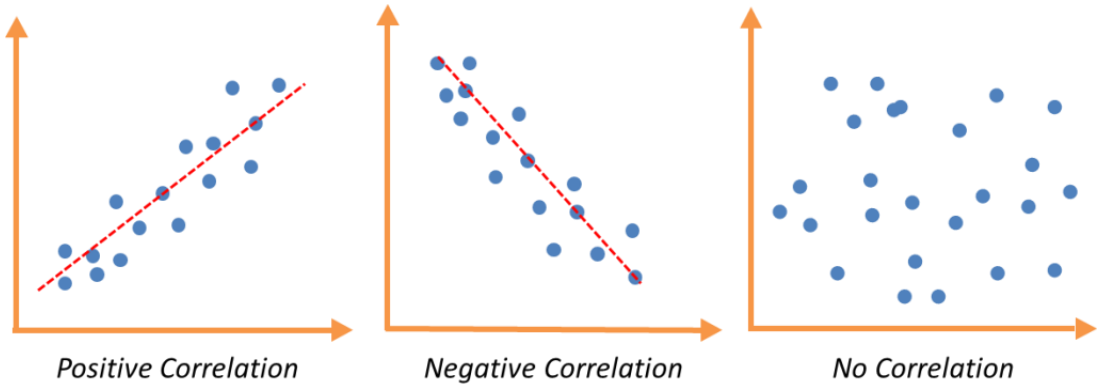
where,

+1 indicates a perfect positive linear relationship,

0 indicates no linear relationship,

-1 indicates a perfect negative linear relationship.

Graphically, a positive correlation in general will display a line of best fit that slopes upwards, a negative correlation will typically show a line of best fit that slopes downwards and data with no correlation will appear scattered with no discernible pattern or trend [7].



Interpretation of Pearson's r can be illustrated as:

Value of r	Interpretation	Example
+0.9 to +1	Very strong positive	Height vs. Weight
+0.6 to +0.9	Strong positive	Study time vs. Exam score
+0.3 to +0.6	Moderate positive	Temperature vs. Ice cream sales
-0.3 to +0.3	Weak or no correlation	Number of Rainy Days vs. Stock Market Returns
-0.6 to -0.3	Moderate negative	Stress Levels vs. Work Productivity
-0.9 to -0.6	Strong negative	Car Speed vs. Fuel Efficiency
-1 to -0.9	Very strong negative	Demand vs. Price

Pearson's correlation need to meet some assumptions to provide reliable insights, such as:

- **Linearity**: The relationship between the variables needs to be linear. Nonlinear associations like exponential or quadratic cannot be presented effectively.
- **Normality**: Both variables should be roughly normally distributed, especially when dealing with small sample sizes.
- **Homoscedasticity**: The variability in one variable should remain consistent across all levels of the other variable.
- **Continuous Data**: The variables need to be measured on an interval or ratio scale.
- **Independence**: Observations should be independent and without autocorrelation as in time-series data.
- **Absence of Outliers**: Extreme values can warp the correlation coefficient and lead to misleading results

3. Applications of Pearson's coefficient in Machine learning

Machine learning, as a subset of artificial intelligence, facilitates computers to learn from experience by adjusting their behavior based on the data they are exposed to. It involves developing algorithms so as to analyze data, identify patterns, and make predictions or decisions [8]. The use of Pearson's correlation coefficient is widely reflected in the following processes of Machine Learning:

3.1 Feature Selection

Feature selection is a vital step in building effective machine learning models. In selecting relevant features, Pearson's correlation helps by showcasing how strongly each feature is linearly allied to the target variable [9]. Features with high positive or negative correlation (closer to +1 or -1) usually exhibit a stronger linear relationship and are more predictive. On the contrary, features with very low correlation (close to 0) are likely to have less useful information for certain models, particularly linear ones.

3.2 Multicollinearity Detection

Multicollinearity represents a situation in which two or more independent variables are highly correlated. High multicollinearity has the potential to distort the training process of linear models, which may lead to unstable coefficients and flimsy generalization. Pearson's correlation can help find out multicollinearity and improve model performance by identifying and removing or assimilating the features.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in predictive analytics and machine learning that entails summarizing, visualizing, and understanding data before modeling. Correlation coefficient, 'r' plays a key role in EDA by finding out linear relationships between variables. This, in turn, facilitates feature selection, multicollinearity detection, and advancing model performance.

3.4 Model Interpretability

Model interpretability in machine learning relates to the understanding of the dynamics of a model's predictions or decisions. It involves comprehending the mechanisms by which the model works and arrives at a particular output. Interpretability helps in finding out errors or biases in the model's training data or logic and thereby allows model improvement and reaching more accurate results. Pearson's correlation facilitates model interpretability by helping to detect how closely model predictions align with actual outcomes, more so in regression tasks. It, thus, assists in building trust, ensuring transparency and removing potential biases in the model, be it a Linear Regression Model, Tree-Based Model or Neural Network.

4. Limitations of Pearson's correlation

Pearson's correlation, despite its importance and extensive usage in predictive analytics and machine learning, has several constrains. Some of the limitations of Pearson's correlation are as follows:

4.1 Linearity assumption:

Pearson's method assumes a linear relationship between variables [10]. It misses complex nonlinear relationships which makes it unsuitable for complex data sets. In cases where the correlation is not well-represented by a straight line, it should be supplemented with mutual information or maximal information coefficient or other alternative correlation measures.

4.2 Outlier sensitivity

Pearson's correlation coefficient is quite sensitive to outliers. Outliers, or data points significantly distant from the main cluster, can disproportionately influence the correlation value [11]. This may lead to a deceptive assessment of the strength and direction of the relationship between variables. Outliers can be eliminated with the application of Spearman's rank correlation or robust scaling.

4.3 Causation ambiguity:

A correlation between two variables doesn't mean one causes the other [10]. The variables can show a strong correlation not because one affects the other, but due to a shared link with a third variable or merely by chance. Inability to distinguish between correlation and causation may lead to faulty conclusions and ineffective interventions. Pearson's correlation, as such, should never be interpreted as substantiation of a causal connection. In fact, it's imperative to ascertain causality through controlled experiments or methods like Granger causality methods, rather than relying solely on observed correlations

4.4 Scope limitation:

The scope of Pearson correlation is confined to bivariate analysis. It can only evaluate the relationship between two variables at a time. While this can be useful for understanding simple, direct relationships, it offers a limited perspective when dealing with complex systems involving multiple interrelated factors. In circumstances where multiple variables are involved alternative methods like multiple regression analysis may provide a more rigorous understanding of the interplay different predictors.

4.5 Scale dependence:

Pearson's coefficient is not robust to changes in scale or measurement units. It is scale-invariant in the sense that linear transformations (such as converting from inches to centimeters) do not change its value, it is still not robust to outliers or to the ordinal nature of data. It assumes interval or ratio data and becomes unreliable with ordinal or skewed data [12]. It is, thus, it is essential to perform thorough data preprocessing before calculating the Pearson correlation. To address the limitation of linear dependence and improve robustness, monotonic rank-based correlation measures like Kendall's Tau or Spearman's Rho can be used as alternatives.

5. Alternative approaches to consider

The Pearson's correlation coefficient is useful for several things, but it does have shortcomings. The limitations set by Pearson's correlation need to be addressed carefully to arrive at an effective estimation. It is thus crucial to recognizing the limitations set by Pearson's correlation coefficient and explore alternative approaches that complement or mitigate the drawbacks. Some alternatives to consider while applying Pearson's correlation are discussed below:

5.1 Spearman's Rank Correlation

Spearman's Rank Correlation assesses monotonic relationships and is less sensitive to outliers. It is predominantly useful when variables are not normally distributed or when the relationship is non-linear but monotonic [13]. It assesses how well the relationship between two variables can be described using a monotonic function.

5.1.1 Spearman's ρ (Without Ties)

If there are no tied ranks, Spearman's ρ is calculated as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (3)$$

Where,

ρ= Spearman's rank correlation coefficient

d_i = difference between ranks of corresponding values($\text{rank}(X_i) - \text{rank}(Y_i)$)

n = number of observations

5.1.2 Spearman's ρ (With Ties)

If tied ranks exist (e.g., same values in *X* or Y), use Pearson's correlation on ranks

$$\rho = \frac{\text{Cov}(\text{rank(x)}, \text{rank(y)})}{\sigma_{\text{rank(x)}}\sigma_{\text{rank(Y)}}} \qquad (4)$$

where, $\text{Cov}(\text{rank(x)}, \text{rank(y)})$ = Covariance of the ranks

$\sigma_{\text{rank(x)}}\sigma_{\text{rank(Y)}}$ = Standard deviations of the ranks

5.2 Standard Kendall's Tau (τ)

Kendall's Tau (τ) is a nonparametric rank correlation coefficient that measures the ordinal association between two variables. Much like Spearman's rank correlation, Kendall's Tau evaluates the strength and direction of a relationship based on the ranks of the data rather than their actual value. What sets Kendall's Tau apart is its approach to handling the ordering of data pairs. Kendall's Tau works well for monotonic relationships (whether linear or nonlinear) and is particularly useful when dealing with small sample sizes. There are two main versions of Kendall's Tau measures:

5.2.1 Kendall's Tau-a (Without Ties)

If there are no tied ranks, Kendall's τ_is calculated as:

$$\tau = \frac{(\text{Number of Concordant Pairs}) - (\text{Number of Discordant Pairs})}{\frac{n(n-1)}{2}} \qquad (5)$$

where,

Concordant Pair: Pairs (x_i,y_i) and (x_j,y_j) where $x_i>x_j$ and $y_i>y_j$; or $x_i<x_j$ *and* $y_i<y_j$

Discordant Pair: Pairs where $x_i>x_j$ and $y_i<y_j$; or $x_i<x_j$ and $y_i>y_j$

Total Pairs: n(n−1)/2 (all possible pairs).

5.2.2 Kendall's Tau-b (τ_e) (Adjusts for Ties)

If tied ranks exist (e.g., $x_i=x_j$ or $y_i=y_j$), use Tau-b:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_x)(n_c + n_d + n_y)}} \qquad (6)$$

where,

n_c = Number of concordant pairs.

n_d = Number of discordant pairs.
t_x = Number of ties in X
t_y = Number of ties in Y.

5.3 Mutual Information

Mutual Information (MI) is a measure from information theory that quantifies the amount of information obtained about one random variable through observing another random variable. MI can detect any statistical dependency, including non-linear and non-monotonic relationships. It is Effective for feature selection in high-dimensional data. It is widely used in feature selection for classification problems and is not constrained by assumptions of normality or linearity

5.4 Distance Correlation

Distance correlation is a newer technique that can detect both linear and non-linear associations. It is increasingly used in high-dimensional machine learning settings where traditional correlation measures fail to capture complex dependencies. It ranges from 0 (no dependence) to 1 (perfect dependence). It can detect complex, non-linear relationships**.**

6. Conclusion

Pearson's correlation coefficient remains a valuable and widely applied statistical tool, particularly for linear relationships between variables. It serves as initial feature screening, thereby reducing dimensionality and model complexity in predictive analytics and machine learning. It aids in detecting multicollinearity among features, which is crucial for model stability and interpretability. Variables demonstrating high linear correlation with the target often merit prioritized data cleaning or imputation efforts, whereas features exhibiting minimal association can be strong initial candidates for exclusion, simplifying subsequent analysis. However, the effective application of Pearson's correlation demands a critical understanding of its inherent limitations. Its primary constraint is its exclusive focus on linear associations; it can completely miss strong non-linear relationships, potentially leading to the erroneous dismissal of useful predictors. Its sensitivity to outliers is another significant weakness, as a few extreme data points can drastically distort the correlation value, providing misleading insights which restricts its standalone use in modern data science tasks. While Pearson's correlation serves as a valuable diagnostic and preprocessing tool in predictive modeling, its limitations necessitate the use of complementary methods. A robust analytical framework should integrate Pearson's correlation with other statistical measures and machine learning techniques to ensure more accurate and meaningful insights from data. By understanding both its capabilities and constraints, data scientists can make more informed choices about when and how to apply these metric, ensuring more accurate and interpretable models.

References

1. V. K. Verma and S. Verma, "Machine learning applications in healthcare sector: An overview," *Materials Today Proceedings*, vol. 57, pp. 2144–2147, Dec. 2021, doi: 10.1016/j.matpr.2021.12.101.
2. M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp.

58–73, Jan. 2022, doi: 10.1016/j.ijin.2022.05.002.
3. D. Broby, "The use of predictive analytics in finance," *The Journal of Finance and Data Science*, vol. 8, pp. 145–161, May 2022, doi: 10.1016/j.jfds.2022.05.003.
4. P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, Feb. 2018, doi: 10.1213/ane.0000000000002864.
5. S. Senthilnathan, "Usefulness of correlation analysis," *SSRN Electronic Journal*, Jan. 2019, doi: 10.2139/ssrn.3416918.
6. Pearson's Correlation Coefficient. In: Kirch, W. (eds) Encyclopedia of Public Health. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569
7. 4. R. J. Janse *et al.*, "Conducting correlation analysis: important limitations and pitfalls," *Clinical Kidney Journal*, vol. 14, no. 11, pp. 2332–2337, May 2021, doi: 10.1093/ckj/sfab085.
8. J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *Journal of Physics Conference Series*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.
9. H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457–5474, Aug. 2021, doi: 10.1007/s10489-021-02524-x.
10. R. J. Janse, T. Hoekstra, K. J. Jager, C. Zoccali, G. Tripepi, F. W. Dekker, and M. van Diepen, "Conducting correlation analysis: important limitations and pitfalls," *Clinical Kidney Journal*, vol. 14, no. 10, pp. 2332–2337, Oct. 2021, doi: 10.1093/ckj/sfab085.
11. A. J. Bishara and J. B. Hittner, "Reducing bias and error in the correlation coefficient due to nonnormality," *Educational and Psychological Measurement*, vol. 75, no. 5, pp. 785–804, Nov. 2014, doi: 10.1177/0013164414557639.
12. R. M. O'Brien, "The Use of Pearson's with Ordinal Data," *American Sociological Review*, vol. 44, no. 5, p. 851, Oct. 1979, doi: 10.2307/2094532.
13. A. J. Bishara and J. B. Hittner, "Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches.," *Psychological Methods*, vol. 17, no. 3, pp. 399–417, May 2012, doi: 10.1037/a0028087.