Heuristics Prediction of Olympic Medals using Machine Learning

Chandrasegar Thirumalai, IEEE Member,
School of Information Technology and Engineering,
VIT University, Vellore, India.
chandru01@gmail.com

Monica S
School of Information Technology and Engineering,
VIT University, Vellore, India.
monicasankar.dpi@gmail.com

Vijayalakshmi A
School of Information Technology and Engineering,
VIT University, Vellore, India.
vijia.2496@gmail.com

***Abstract*— This paper determines methods to develop a novel technique for predicting a nation in view of the Olympic awards owned by 2012. It is the combination of three methods are Pearson correlation coefficient, Spearman correlation coefficient and along with linear regression. The main idea of the paper is to compare the value of Spearman and Pearson correlation coefficient as there in the same set of data. The example concerns the comparison of the total medals and the GDP (gross domestic product) that has been obtained by each country. The results from using these methods do the heuristics prediction of Olympic medals using machine learning.**

Keyword: Pearson method, Spearman method, Linear Regression, Number of the static attributes (NSF).

I. INTRODUCTION

The 2012 Summer Olympics, prominently known as London 2012, occurred in London and different places over the United Kingdom from 25 July to 12 August 2012. A sum of 10,768 competitors from 204 National Olympic Committees (NOCs) took an interest in the super occasion. London 2012 Olympics Games award is being touted as the greatest Summer Olympics decorations to date [1].

The informational index gives the names of the taking part nations and the aggregate number of decorations won by the nation. The second properties depict about the aggregate awards which incorporate a number of gold, silver and bronze decorations won by nation, and taking the third qualities called total national output (GDP= wage per capita increased by populace estimate).

Every gold award is comprised of 92.5 percent silver and 1.34 percent gold, with the rest of. The silver decoration (which speaks to second place) is comprised of 92.5 percent silver, with the rest of. The bronze decoration is comprised of 97 percent Copper, 2.5 percent Zinc, and 0.5 percent Tin [2]. Here we have taken the aggregate awards of the nations in light of which have high, medium and low of five GDP esteem in every classification. The information is publically accessible and was acquired from et. al [3], [39].

TABLE 1. COUNTRY, TOTAL MEDAL, AND GDP

Country	*Total Medal*	*GDP*
United States	104	1.519467
China	88	0.725389
Japan	38	0.586871
Germany	44	0.357877
France	34	0.279577
Cyprus	1	0.00336
Libya	0	0.003228
Ethiopia	7	0.003035
Jordan	0	0.003026
Panama	0	0.0029
Nauru	0	2.34E-06
Tuvalu	0	3.58E-06
Marshall Island	0	1.15E-05
Kiribati	0	1.65E-05
Palau	0	1.70E-05

This is the original dataset which is publicly available in the [3]. Here the value represented Olympic dataset attributes are sorted by three which are a high correlation that range is equal to (-0.5 t0 1), medium correlation is equal to (-0.3 to 0.5) and low correlation is equal to (-0.1 to 0.3).

We have a dataset portrayed the Olympic decoration as appeared in Table 1. This informational collection is utilized as the contribution to figure the straight relapse [24], [25] and Pearson [4], [8], [12], [14], [20], [22]. In the present days, there are tremendous measures of information recorded by the Olympic council and examining them requires complex calculations. We played out the product metric examination on the given informational index. From the information investigation [6], [7], [10], [13], [16], [18] we can choose

which trait can be considered and which property can be ignored. A portion of the previous techniques to figure the choices in view of their relationship of quality are Spearman [9], Analytical Hierarchical Process (AHP) [5], [11], [13] and Traveling Salesman Problem (TSP) [36]. The touchy data's among different substances [17], [19], [26], [28], [30], [32], [34], [38] among the bank stock model are taken care of by late secured techniques [21], [23], [27], [29], [31], [33], [35], [37].

II. METHODS

A. Pearson method

Correlation is a method for exploring the relationship between two quantitative, constant factors, for instance, country, total decorations and gross domestic product taken in the dataset. Pearson's correlation coefficient (r) is a measure of the nature of the relationship between the two components.

The initial phase in concentrate the relationship between two consistent factors is to draw a diffuse plot of the factors to check for linearity. The connection coefficient ought not to be computed if the relationship is not direct. For correlation just purposes, it doesn't generally make a difference on which hub the factors are plotted. Be that as it may, customarily, the free (or illustrative) variable is plotted on the x-hub (on a level plane) and the ward (or reaction) variable is plotted on the y-hub (vertically).

The closer disperse of focuses is to a straight line, the higher the quality of the relationship between the factors. Additionally, it doesn't make a difference what estimation units are utilized.

$$r=\frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{[n\sum x^2-(\sum x)^2[N\sum Y^2-(\sum Y)^2]}}$$

B. Spearman method

Spearman correlation coefficient evaluates the linear relationship between two continuous variables. And it is based on the ranked values for each variable rather than the raw data.

Spearman's Rank connection coefficient is a method which can be utilized to bridge the quality and bearing (negative or positive) of a relationship between two factors.

1. Make a relation between your information.
2. Rank the two informational indexes. Positioning is accomplished by giving the positioning "1" to the greatest number in a segment, "2" to the second greatest esteem et cetera. The littlest incentive in the section will get the least positioning. This ought to be accomplished for both arrangements of estimations.
3. Tied scores are given the mean (normal) rank
4. Discover the distinction in the positions (d): This is the contrast between the positions of the two values on each column of table 1. The rank of the second esteem (add up to decorations) is subtracted from the rank of the main (GDP).
5. Square the distinctions (d^2). To evacuate negative values and after that whole them (d^2)

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

C. Linear Regression

Straight Regression is an expectation when a variable (y) is reliant on the second variable(x) in light of the relapse condition of a given arrangement of information. Linear relapse endeavors to demonstrate the relationship between two variables by fitting the direct condition to watched information. One variable is thought to be a needy variable.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{n}(h_\theta(x_i) - y_i)^2$$

III. NUMERICAL ANALYSIS

A. Pearson correlation coefficient:

TABLE 2. PEARSON CORRELATION CALCULATED VALUE

N	15
∑XY	111.412
∑X	212
∑Y	1.967
∑X^2	12330
∑Y^2	1.0769
1254.2	
2E+06	1311.45

The outcomes will be between - 1 and 1. You will seldom observe 0, - 1 or 1. You'll get a number some place in the middle of those qualities. The nearer the estimation of r gets the chance to zero, the more prominent the variety the information focuses are around the line of best fit, r = 0.95. In Pearson correlation coefficient the calculate r is 0.95. This value ranges between 0.5 to 1.0 which leads to the high correlation. So the correlation between total medals and GDP is high.

B. Spearman correlation coefficient

This general formula for spearman correlation coefficient. Using this formula, we calculated the correlation method with the help Olympic dataset. The calculated values are here,

TABLE 3. SPEARMAN CORRELATION CALCULATED VALUE

$\sum di^2$	163
$6\sum di^2$	978
$n(n^2-1)$	3360
$\frac{6\sum di^2}{n(n^2-1)}$	0.291071
$1-\frac{6\sum di^2}{n(n^2-1)}$	0.708929

1. On the off chance that it is beneath the line stamped 5%, then it is conceivable your outcome was the result of shot and you should dismiss the theory.

2. In the event that it is over the 0.1% criticalness level, then we can be 99.9% certain the relationship has not happened by the possibility.

3. On the off chance that it is over 1%, yet beneath 0.1%, you can state you are 99% sure.

4.On the off chance that it is over 5%, yet underneath 1%, you can state you are 95% sure (i.e. measurable there is a 5% probability the outcome happened by shot).

In the Spearman correlation coefficient, the calculate r is 0.70.

This value ranges between 0.5 to 1.0 which leads to the high correlation. So the correlation between total medals and GDP is high

C. Linear regression

TABLE 4. COST OPTIMIZATION USING LINEAR REGRESSION

θ	*0*	*0.5*	*1*	*1.5*	*2*
X	23146	22877.41	22610.51	22345.31	22081.8
J(θ)	771.53	762.58	753.68	744.84	736.05

Here θ=1.5 takes the optimized cost on Y.

The graph will predict the total medal and GDP value will the sorted country attributes.

X axis which represents the GDP value.
Y axis which represents the total medal value.

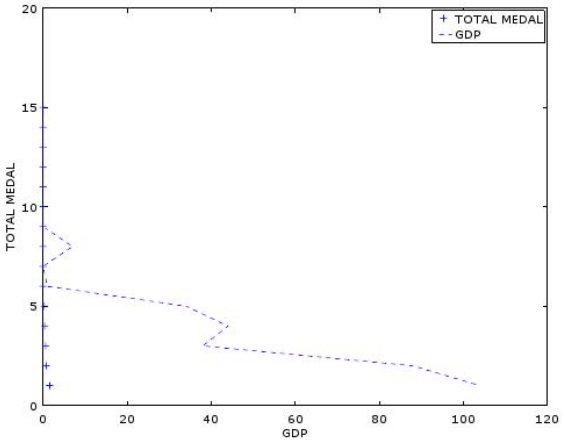

Figure 1. GDP, Total Medal, and Country.

IV. CONCLUSION

Both the spearman and Pearson gives high correlation. But the calculated value of spearman is less the Pearson correlation value. In the given Olympic dataset we predict the GDP (gross domestic product) values and it will predict the value along with the country. So united states as the highest total medal value and predicted the value of r=0.95. So, the united states as the highest possibility to own the next Olympic proved by the Pearson correlation method. In conclusion, correlation method played a major role in predicts the gross domestic product at the London 2012 summer Olympic.

REFERENCES

[1] http://olympics.sporting99.com/london-2012/index.html

[2] DeMarco, Anthony (26 July 2012). "London's Olympic Gold Medal Worth The Most In The History Of The Games". Forbes. Retrieved 30 July 2012

[3] http://espn.go.com/olympics/summer2012/decorations and the rundown of countries at http://www.london2012.com/nations/

[4] Hauke J., Kossowski T., Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Quaestiones Geographicae 30(2), Bogucki Wydawnictwo Naukowe, Poznań 2011, pp. 87–93, 3

[5] Piovani J.I., 2008. The historical construction of correlation as a conceptual and operative instrument for empirical research. Quality & Quantity 42: 757–777.

[6] P. Dhavachelvan, Chandra Segar T, K. Satheskumar, "Evaluation of SOA Complexity Metrics Using Weyuker's Axioms," IEEE International Advance Computing (IACC), India, pp. 2325 – 2329, March 2009

[7] Halstead Metric for Intelligence, Effort, Time predictions, DOI:10.13140/RG.2.2.17988.42881

[8] Fisher R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1: 3–32.

[9] Spearman C.E, 1904b. General intelligence objectively determined and measured. American Journal of Psychology 15: 201–293.

[10] Software metric Numerical Data analysis using Box plot and control chart methods, VIT University, DOI:10.13140/RG.2.2.27422.95041

[11] Vaishnavi B, Karthikeyan J, Kiran Yarrakula, Chandrasegar Thirumalai, "An Assessment Framework for Precipitation Decision Making Using AHP", International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[12] Griffith D.A., 2003. Spatial autocorrelation and spatial filtering. Springer, Berlin.

[13] Chandrasegar Thirumalai, Senthilkumar M, "An Assessment Framework of Intuitionistic Fuzzy Network for C2B Decision Making", International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[14] Rodgers J.L. & Nicewander W.A., 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1): 59–66.

[15] F. Fioravanti, P. Nesi, "A method and tool for assessing object-oriented projects and metrics management," Journal of Systems and Software, Volume 53, Issue 2, 31 August 2000, Pages 111-136

[16] Galton F., 1875. Statistics by intercomparison. Philosophical Magazine 49: 33–46

[17] Chandrasegar Thirumalai, Viswanathan P, "Diophantine based Asymmetric Cryptomata for Cloud Confidentiality and Blind Signature applications," JISA, Elsevier, 2017.

[18] Galton F., 1877. Typical laws of heredity. Proceedings of the Royal Institution 8: 282–301.

[19] Chandrasegar Thirumalai, Sathish Shanmugam, "Multi-key distribution scheme using Diophantine form for secure IoT communications," IEEE IPACT 2017.

[20] Galton F., 1888. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London 45: 135–145.

[21] Chandrasegar Thirumalai, Senthilkumar M, "Spanning Tree approach for Error Detection and Correction," IJPT, Vol. 8, Issue No. 4, Dec-2016, pp. 5009-5020.

[22] Galton F., 1890. Kinship and correlation. North American Review 150: 419–431.

[23] Chandrasegar Thirumalai, Senthilkumar M, "Secured E-Mail System using Base 128 Encoding Scheme," International journal of pharmacy and technology, Vol. 8 Issue 4, Dec. 2016, pp. 21797-21806.

[24] Yule G.U., 1897a. On the significance of Bravais' formulae for regression, in the case of skew correlation. Proceedings of the Royal Society of London Ser. A 60: 477–489

[25] Chandramowliswaran N, Srinivasan.S and Chandra Segar.T, "A Note on Linear based Set Associative Cache address System" International J. on Computer Science and Engg. (IJCSE) & India, Engineering Journals & 0975-3397, Vol. 4 No. 08 / pp. 1383-1386 / Aug. 2012.

[26] T Chandra Segar, R Vijayaragavan, "Pell's RSA key generation and its security analysis," in Computing, Communications and Networking Technologies (ICCCNT) 2013, pp. 1-5

[27] Chandrasegar Thirumalai, Senthilkumar M, Vaishnavi B, "Physicians Medicament using Linear Public Key Crypto System," in International conference on Electrical, Electronics, and Optimization Techniques, ICEEOT, IEEE & 978-1-4673-9939-5, March 2016.

[28] Chandrasegar Thirumalai, "Physicians Drug encoding system using an Efficient and Secured Linear Public Key Cryptosystem (ESLPKC)," International journal of pharmacy and technology, Vol. 8 Issue 3, Sep. 2016, pp. 16296-16303

[29] E Malathy, Chandra Segar Thirumalai, "Review on non-linear set associative cache design," IJPT, Dec-2016, Vol. 8, Issue No.4, pp. 5320-5330

[30] "DDoS: Survey Of Traceback Methods", International Joint Journal Conference in Engineering 2009, ISSN 1797-9617.

[31] Chandrasegar Thirumalai, Senthilkumar M, Silambarasan R, Carlos Becker Westphall, "Analyzing the strength of Pell's RSA," IJPT, Vol. 8 Issue 4, Dec. 2016, pp. 21869-21874.

[32] Chandramowliswaran N, Srinivasan.S and Chandra Segar T, "A Novel scheme for Secured Associative Mapping" The International J. of Computer Science and Applications (TIJCSA) & India, TIJCSA Publishers & 2278-1080, Vol. 1, No 5 / pp. 1-7 / July 2012

[33] Chandrasegar Thirumalai, "Review on the memory efficient RSA variants," International Journal of Pharmacy and Technology, Vol. 8 Issue 4, Dec. 2016, pp. 4907-4916.

[34] Vinothini S, Chandra Segar Thirumalai, Vijayaragavan R, Senthil Kumar M, "A Cubic based Set Associative Cache encoded mapping," International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 02 May -2015

[35] Chandrasegar Thirumalai, Himanshu Kar, "Memory Efficient Multi Key (MEMK) generation scheme for secure transportation of sensitive data over Cloud and IoT devices," IEEE IPACT 2017.

[36] M.Senthilkumar, T.Chandrasegar, M.K. Nallakaruppan, S.Prasanna, "A Modified and Efficient Genetic Algorithm to Address a Travelling Salesman Problem," in International Journal of Applied Engineering Research, Vol. 9 No. 10, 2014, pp. 1279-1288

[37] Nallakaruppan, M.K., Senthil Kumar, M., Chandrasegar, T., Suraj, K.A., Magesh, G., "Accident avoidance in railway tracks using Adhoc wireless networks," 2014, IJAER, 9 (21), pp. 9551-9556.

[38] Chandrasegar Thirumalai, Viswanathan P, "Hybrid IT architecture with Gene based Cryptomata (HITAGC) for mutual authentication and privacy preserving security services," International Journal of Advanced Intelligence Paradigms, 2017.

[39] Young, A. W.; Cave, B. M.; Lee, A.; Pearson, K. (1917). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher.

S Monica currently pursuing M.S Software Engineering at VIT University, Vellore Campus, Vellore, India.

A Vijayalakshmi currently pursuing M.S Software Engineering at VIT University, Vellore Campus, Vellore, India.