IMPROVED ACCURACY IN PREDICTING THE NETWORK PERFORMANCE FOR ALLOCATING SERVER RESOURCES USING LINEAR REGRESSION COMPARED WITH RANDOM FOREST.

Gade. Mary Spandana[1] , K. SashiRekha[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The major goal of this study is to evaluate the accuracy of the Novel Linear Regression technique for forecasting network performance for allocating server resources to the Random Forest (RF) approach utilizing meteorological data.

Materials and Methods: The Novel Linear Regression algorithm (LR) with sample size=10, 95 percent confidence interval, and pretest power of 80 % was iterated at various times to predict the accuracy percentage of network performance for allocating server resources in technology infrastructure. Novel Linear Regression allows for more accuracy by transforming data into a higher-dimensional space.

Results: Novel Linear Regression appears to prove with better accuracy (88%) compared to Random Forest accuracy (74%). There was an insignificant difference between LR and RF with p=0.818 ($p<0.05$).

Conclusion: For estimating network performance, Novel Linear Regression performed with better accuracy compared to Random Forest.

Keywords: Novel Linear Regression, Random Forest, Dependent, Performance, Accuracy, Machine Learning, Testing, Training.

[1]Research Scholar, Department of Computer Science Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

[2*]Project Guide, Department of Computer Science Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

1. Introduction

Novel Linear Regression appears to be more effective than Random Forest at identifying network performance. Network Performance and Security: Testing and Analyzing Using Open Source and Low-Cost Tools (Chapman 2016). A Linux Server Operating System's Performance Comparison Using Lmbench (Jiang 2016). Performance of Multi Server Validation and Key Association with User Protection in Network Security (Padmini et al. 2016). Performance Analysis and Optimization of Network Server Based on Queuing Network (Zhang, Jing-tai, and Wang 2011). The applications of the research are students will be able to demonstrate digital literacy through hardware, operating systems, networking etc. Applications of Machine learning to resource management in cloud computing (Huang et al. 2013).

In order to bring awareness to network performance in server resources, many researchers have adopted identifying network performance for allocating server resources. "Phaeton: A Log-Based Architecture for High Performance File Server Design." Proceedings Pacific Rim International Symposium on Fault-Tolerant Systems (Chin, C., Chung-Kie Tung, and Shang-Rong Tsai). There are around 1700 papers in Google Scholar. Around 120 publications have been published by IEEE. Web Server Performance Enhancement by Suppressing Network Traffic for High Performance Client (Jin and Tomoishi 2015). In network performance, there are a range of methods for allocating server resources in IT architecture. With the biggest variance score (R2) of 0.85260 and the smallest bias score (RMSE) of 0.02323, Novel linear regression modeling had a respectable score. (Wootton, C. 2007) "What Your Web Server Log Can Tell You." Developing Quality Metadata. With kNN modeling, a decent result was attained, with a variation score of 0.57422 and a bias score of 0.08049. The highest referenced study (Heruwidagdo et al. 2021) focused on estimating network performance of server resources using Random Forest, which had an accuracy of 72%. Although many people may not respond to the data and may not always provide reliable information, collecting ground-truth data may not yield absolute accuracy "Measuring the Network Performance of Google Cloud Platform." (Mok et al. 2021). According to the literature review, machine learning algorithms have been used to forecast network performance of server resources with high accuracy. Network and Server Resource Management Strategies for Data Center Infrastructures (Tso, Fung Po, Simon Jouet, and Dimitrios P. Pezaros 2020). The Random Forest approach is commonly used to handle classification and regression problems.Our team has extensive knowledge and research experience that has translated into high quality publications (Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa et al. 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; Mohan et al. 2022)

The research gap identified for the existing work is that there are some failures in the web server crashes in prediction of network performance for allocating the server resources. Machine learning can be used to teach how to predict the network performance for servers. This proposed system is to improve accuracy in predicting the network performance in server resources.

2. Materials and Methods

The analysis was carried out in the Open Source Lab of Saveetha School of Engineering of the Saveetha Institute of Medical and Technical Sciences (SIMATS). There will be two groups in this research. The first group contains the Novel Linear Regression algorithm, whereas the second contains the Random Forest algorithm. Extracting usage patterns from web server log (Jeba, J. Monisha Privthy, J. Monisha Privthy Jeba, M. S. Bhuvaneswari, and K. Muneeswaran. 2016). Each group's sample size is determined using data from a previous study with an 80% g power. Based on available data, the sample size for the Novel Linear Regression method (N=10) and the Random Forest algorithm (N=10) were calculated Predicting Utilization of Server Resources from Log Data (Vora, Mehul Nalin. 2020). To begin, import the dataset into Collab (Jeba et al. 2016). Following the import of the dataset, the relevant code for measuring the results of the Novel Linear Regression was provided. The code's desired output is later acquired in the form of graphs and values that have been provided.

Linear Regression

In the domains of statistics and machine learning, Novel linear regression is one of the most well-known and so well methods. Novel Linear regression is a statistical technique for determining the connection here between scalar response and one or more independent variables (also known as dependent and independent variables). Simple Novel linear regression is used when there is only one mediating variable, and multiple linear regression has been used when there are several mediating variables.

Algorithm

1. Start Program
2. Input the dataset
3. Explore the data to figure out what they look like
4. Give path to LR to configure and store it in a variable
5. Start while loop to loop frame one by one
6. Detect the objects in frames using variables that we declared initially in step 2
7. find the centroid points of each side(x,y)

7.1)calculating the words

7.2) Pre-process the data

7.3)Remove nan values and unwanted values

7.4)clearing the stop words using lemitizing

7.5)Removing the unwanted text by cleaning and stemming process

8. Split the data into attributes and tables
9. Divide the data into training and testing sets
10. Train the K-Nearest neighbor algorithm
11. Destroy all windows
12. End program.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv(' mac server dataset.csv ')
X = dataset.iloc[: , :-1].values
Y = dataset,iloc[: , 1].values
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X , y , test_size =%, random_state = 1)
from sklearn.linear_model import LinearRegression
regression = LinearRegression()
regressor.fit(X_train , y_train)
y_pred = regressor . predict (X_test)
plt.scatter (X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict (X_train), color = 'blue')
plt.scatter (X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title(' Salary vs Experience (Test set) ' )
plt.xlabel(' Years of Experience')
plt.ylabel ('Salary')
plt.show()
```

Fig. 1. Pseudocode of Linear Regression Algorithm

Random Forest

A random forest is a machine learning approach for solving classification and regression problems. It makes use of supervised methods, which is an approach for solving complicated issues by combining several classifiers. Many decision trees make up a random forest algorithm. Both classification and regression tasks can benefit from the Random Forest technique. Cross validation ensures a higher level of accuracy. The missing values will be handled by the random forest classifier, which will retain the correctness of a major percentage of the data.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv(' mac server dataset.csv ')
x = dataset.iloc[:,[1,3]].values
y = dataset.iloc[:,4].values
From sklearn.model_selection import
train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25, random_state = 0)
from sklearn.preprocessing import StandardScalar
sc = StandardScalar
x_train - sc.fit_transform(x_train)
x_test = sc.transform(x_test)
import RandomForestClassifier
classifer= RandomForestClassifier(n_estimators =10, criterion = 'entropy',random_state =0)
classifer.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
cm = confusion_matrix(y_test, y_pred)
percentage = score*100
print(f Accuracy: {round(percentage,3)}%')
```

Fig. 2. Pseudocode of Random Forest Algorithm

Testing Procedure

The dataset was split into two parts: one for training and one for testing. Each algorithm was already trained based on the best attribute from the training phase. Testing will be employed after training to compare the taught set with the output, which will then be provided as an accuracy rate. After that, we implemented the code for the Random Forest methodology to extract the values for some data. Using Google Collab and the Python programming language, the Novel Linear Regression Algorithm and Random Forest were examined. An Intel Core i5 processor and 8GB of RAM were included in the system configuration. A 64-bit operating system, an X64-based processor, and a 917-GB hard drive were all included in the workstation. The software configuration includes the Windows 10 operating system. For the analysis, IBM SPSS (Statistical Package for Social Sciences) version 21 was employed. It's a form of statistical programme for analyzing data. Ten iterations with a maximum of 20 samples were performed for both proposed and present algorithms, and the projected accuracy for each iteration was recorded for accuracy analysis.

Statistical Analysis

The Statistical Package for the Social Sciences(SPSS) was used to conduct the statistical analysis (Ebbers et al. 2013) and Google collab software tools. Descriptive statistics for mean, standard deviation, and relative deviation were obtained for the LR and Decision Tree algorithms. Independent variables are anonymized logging parameters X 1, X 2, and so on. The dependent variable was output variables (Accuracy). An independent sample t-test is used to compare the performance of algorithms.

3. Results

In this investigation, ten samples were tested for Novel Linear Regression and Random Forest, with 88 percent and 74% accuracy, respectively, as shown in Table 1, with different occurrences occuring on different days.

Table 1 also shows the mean, standard deviation, and standard error mean of the Novel Linear Regression and Random Forest algorithms for network performance to server resources, revealing that the Novel Linear Regression algorithm has an accuracy mean of 85 percent, standard deviation 2.95419 for sample size of N=10, and the Random Forest algorithm after predicting has an accuracy mean of 69 percent, standard deviation 3.12358 for sample size of N=10. The mean, standard deviation, and significant difference between Novel Linear Regression and Random Forest suggest that there is a substantial difference between the two groups. The bar graph Fig. 1 displays the average accuracy of the Novel Linear Regression algorithm and the Random Forest technique for network performance of server resources.

4. Discussion

In terms of accuracy, the Novel Linear Regression method trumps the Random Forest strategy, according to the findings. The standard deviation of Novel Linear Regression was 2.95419, while the standard deviation of Random Forest was 3.12358. As indicated in Table 2, the standard error mean for Novel Linear Regression was 0.93420, while the standard error mean for the Random Forest approach was 0.98776, for a sample size of ten people. As a conclusion, when it comes to network performance when distributing server resources, the Novel Linear Regression method helps performance the Random Forest approach. A gpower of 80 percent was used for the pretest analysis. There are numerous techniques used in network performance for allocating server resources in IT infrastructure. The accuracy of the Random Forest algorithm was 81% (Heru Widagdo et al. 2021). In network performance, there are a variety of ways for allocating server resources in information technology (IT) architecture. Novel Linear regression modeling produced a decent performance, with the highest variance score (R2) of 0.85260 and the smallest bias score (RMSE) of 0.02323 (Syme, Matthew, and Philip Goldie. 2004). To demonstrate the difference between the Novel Linear Regression and Random Forest algorithms, a graph is created based on the collected data. Using the accuracy of the algorithms, we were able to achieve good results. RMSE Score Bias for Random Forest 0.036077 and Score Variance 0.644596 (Roy, Chiranjiv, Sourov Moitra, Mainak Das, Subramaniyan Srinivasan, and Rashika Malhotra. 2020). KNN modeling provided an excellent result, with a variation score of 0.57422 and a bias score of 0.08049. Data that has been converted into a tabular format is used to develop training and test materials. “Resource Monitoring on an Application Layer Active Network Server (Liabotis, I., D. Garton, O. Prnjat, and L. Sacks). The data for this study was obtained by extracting and analyzing information from log files retrieved from a server device, and then converting it into a model. The most useful source of information for figuring out how a machine functions is this log. Machine learning-based data analysis is expected to yield a more accurate estimate of server requirements. The

accompanying are the obstacles in accessing and maintaining these logs, according to the following: The wide range of products supplied (servers, storage, networking, backup power, and so on) The number of different goods manufacturers (multi manufacturer), as well as related things from different sellers (Servers, Storage, and Networking), that can be used in a same network. The future work can be implemented in companies to improve the network performance.

5. Conclusion

In this proposed work, prediction of network performance for allocating server resources is performed using Linear Regression algorithm and has improved accuracy 88.00% when compared to the Random Forest (RF) algorithm.

Declaration

Conflict of interests

No conflict of interest in this manuscript.

Authors Contributions

Author GMS was involved in data collection, data analysis, manuscript writing. Author KSR was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Qbec Infosol
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

6. References

Chapman, Chris. 2016. Network Performance and Security: Testing and Analyzing Using Open Source and Low-Cost Tools. Syngress.

Jiang, Zexin. 2016. "A Linux Server Operating System's Performance Comparison Using Lmbench." 2016 International Conference on Network and Information Systems for Computers (ICNISC). https://doi.org/10.1109/icnisc.2016.043.

Jin, Yong, and Masahiko Tomoishi. 2015. "Web Server Performance Enhancement by Suppressing Network Traffic for High Performance Clients." 2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS). https://doi.org/10.1109/apnoms.2015.7275385.

Padmini, R., Sri Padmavati Mahila Visvavidyalayam, School of Engineering and Technology, Padmavati Nagar, Tirupati, Andhra Pradesh, and INDIA. 2016. "Performance of Multi Server Validation and Key Association with User Protection in Network Security." International Journal Of Engineering And Computer Science. https://doi.org/10.18535/ijecs/v5i10.67.

Syme, Matthew, and Philip Goldie. 2004. Optimizing Network Performance with Content Switching: Server, Firewall, and Cache Load Balancing. Prentice Hall Professional.

Zhang, Zheng, L. I. U. Jing-tai, and Hong-Peng Wang. 2011. "Performance Analysis and Optimization of Network Server Based on Queuing Network." Journal of Computer Applications. https://doi.org/10.3724/sp.j.1087.2010.03148.

Tso, Fung Po, Simon Jouet, and Dimitrios P. Pezaros. 2016. "Network and Server Resource Management Strategies for Data Center Infrastructures: A Survey." Computer Networks. https://doi.org/10.7763/ijmo.2013.v3.256.

Ravindran, Kaliappa, Mohammad Rabby, and Shereef El Metwally. 2010. "Management Intelligence for Optimal Resource Allocations in Network Server Systems." 2010 IEEE Network Operations and Management Symposium - NOMS 2010. https://doi.org/10.1109/noms.2010.5488497.

Jin, Yong, and Masahiko Tomoishi. 2015. "Web Server Performance Enhancement by Suppressing Network Traffic for High Performance Clients." 2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS). https://doi.org/10.1109/apnoms.2015.7275385.

Khanna, G., K. Beaty, G. Kar, and A. Kochut. 2006. "Application Performance Management in Virtualized Server Environments." 2006 IEEE/IFIP Network Operations and Management Symposium NOMS 2006. https://doi.org/10.1109/noms.2006.1687567.

Terrell, Jeff, Kevin Jeffay, F. Donelson Smith, Jim Gogan, and Joni Keller. 2008. "Exposing Server Performance to Network Managers through Passive Network Measurements."

2008 IEEE Internet Network Management Workshop (INM).

Han, Jeong Soo, Seong Jin Ahn, and Jin Wook Chung. "Web-Based Performance Manager System for a Web Server." NOMS 98 1998 IEEE Network Operations and Management Symposium. https://doi.org/10.1109/noms.1998.654836.

Liabotis, I., D. Garton, O. Prnjat, and L. Sacks. "Resource Monitoring on an Application Layer Active Network Server." NOMS 2002. IEEE/IFIP Network Operations and Management Symposium." Management Solutions for the New Communications World"(CatNo.02CH37327). https://doi.org/10.1109/noms.2002.1015641.

Zhou, Jin. "Multiple Evidence Combination for Web Site Search Using Server Log Analysis." https://doi.org/10.32920/ryerson.14658057.v1.

Yazdi, M. Amin, M. Amin Yazdi, Pejman Farhadi Ghalatia, and Benedikt Heinrichs. 2021."Event Log Abstraction in Client-Server Applications." Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. https://doi.org/10.5220/001065200000306.

Wootton, C. 2007. "What Your Web Server Log Can Tell You." Developing Quality Metadata. https://doi.org/10.1016/b978-0-240-80869-7.50045-6.

Mok, Ricky K. P., Hongyu Zou, Rui Yang, Tom Koch, Ethan Katz-Bassett, and K. C. Claffy. 2021. "Measuring the Network Performance of Google Cloud Platform." In Proceedings of the 21st ACM Internet Measurement Conference. New York, NY, USA: ACM. https://doi.org/10.1145/3487552.3487862.

Tso, Fung Po, Simon Jouet, and Dimitrios P. Pezaros. 2016. "Network and Server Resource Management Strategies for Data Center Infrastructures: A Survey." Computer Networks 106 (September): 209–25.

Roy, Chiranjiv, Sourov Moitra, Mainak Das, Subramaniyan Srinivasan, and Rashika Malhotra. 2020. "IT Infrastructure Downtime Preemption Using Hybrid Machine Learning and NLP." Annals of Computer Science and Information Systems. https://doi.org/10.15439/2015f400.

Tables and Figures

Table 1:- Accuracy values for Novel Linear Regression and Random Forest.

S No.	Linear Regression (Accuracy)	Random Forest (Accuracy)
1	88.00	74.77
2	87.01	73.01
3	86.02	72.41
4	85.03	71.30
5	84.00	70.51
6	83.23	69.53
7	82.41	68.00
8	81.21	67.51
9	80.00	66.21
10	79.21	65.13

Table 2:- The Novel Linear Regression approach has a mean and standard deviation comparison accuracy of 88%, whereas the Random Forest technique has a mean and standard deviation comparing accuracy of 74%.

S No	N	Mean	Standard Deviation	Std.Mean Error

Linear Regression	10	85.6120	2.95419	.93420
Random Forest	10	69.8380	3.12358	.98776

Table 3: With a 95 percent confidence interval and a significance criterion of 0.01 (the Linear Regression algorithm performs significantly better than the Random Forest approach with a value of 0.818), the result of an Independent T-test is generated.

		Levene's Test for Equality of Variances		**T-test for equality of Means**						
		F	**Sig**	**t**	**df**	**Sig (2-tai-led)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence interval of the Difference**	
									Lower	**Upper**
Accu-racy	**Equal Variances assumed**	.055	.818	11.602	18	.001	15.7740	1.35956	12.9176	18.6303
	Equal variances not assumed			11.602	17.944	.001	15.7740	1.35956	12.9170	18.6309

Fig. 4.

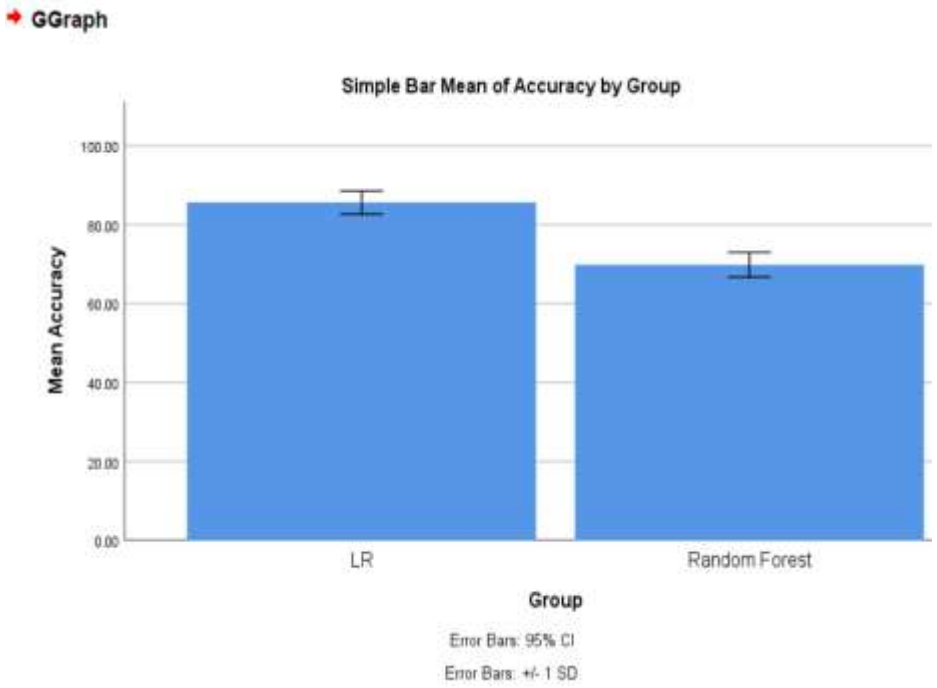

Fig. 3. The Novel Linear Regression algorithm and the Random Forest algorithm are contrasted in terms of mean accuracy. Random Forest has a greater mean accuracy (85%) than Novel Linear Regression (69 percent). The standard deviation of the Novel Linear Regression algorithm is lower than that of the Random Forest approach. On the X Axis, LR vs. Random Forest. Mean detection accuracy 1 SD on the Y axis.