

Forecasting of Cloud Computing Services Workload using Machine Learning

Krishan Kumar¹, K. Gangadhara Rao², K. Gangadhara Rao³, Suneetha Bulla⁴

¹Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, India.

²Professor, Department of CSE, Acharya Nagarjuna University, Guntur, India.

³Associate Professor, Department of CSE, Acharya Nagarjuna University, Guntur, India.

⁴Suneetha Bulla, Associate Professor, Koneru Lakshmaiah Education Foundation

¹suneethabulla@gmail.com,²kancherla123@gmail.com,³bobbababao@gmail.com,⁴suneethabulla@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: This paper analyses and compares prediction accuracy of different machine learning algorithms intended to forecast the workloads of server logs. The proposed prediction model conducts comparative study has been applied using Linear Regression (LR), K- Nearest Neighbors (KNN), Support Vector Machine (SVM), ARMA, ARIMA, and Support Vector Regression (SVR) for web applications to select the suitable algorithm as per workload features. The experiments have used real trace files to evaluate the best suitable method to predict the workloads. The experimental results describe that the ARIMA model shows significant improvement in QoS metrics and improve the cloud datacenter availability in a cloud environment and forecasting. Finally results presented and conclusions are drawn.

1. Introduction

Cloud Computing (CC) is one of the dominant technologies withinside the real time/online packages and has turn out to be one of the quickest developing due to the motive that numerous businesses have migrated from neighborhood computing infrastructure to cloud infrastructure for decreasing the bodily aid expenses which could demand prematurely infrastructure spending. CC has been diagnosed with the aid of using Gartner as one of the pinnacles 10 technology and declared that CC performs an essential position in earnings of businesses [1]. This is an Internet orientated computing wherein cloud assets like software program, hardware infrastructure, platform, gadgets, and internet services are to be had on a version called pay-as-you-go. Cloud customers undertake both hardware and software program digital assets from carrier vendors on price basis as they make use of as an alternative of investing themselves on assets. CC infrastructures offer 3 sorts of services via centralized information facilities and host internet packages [3].

NIST(National Institute of Standards and Technology) described CC as a version for allowing the on-call for, ubiquitous, handy and international community in and out get right of entry to to a distribute pool of configurable computing infrastructure like servers, applications, networks, offerings and storage, which could be provisioned and launched with the minimum provider company intervention or the control effort. Cloud has several capabilities which permit it to serve its clients effectively. Cloud capabilities consist of Scalability, flexibility, on-call for self-provider provisioning and elasticity [2][4].

The arrival rate at cloud datacenters withinside the shape of task [5] sends via way of means of the users. Every task consists of positive self-defining attributes together with the computing time, person authentication, and its respective useful resource necessities in phrases of infrastructure. As single task may also incorporate one or extra responsibilities, which are scheduled for processing on the cloud servers. Tasks also are sure to have numerous carrier necessities together with throughput, latency, and jitter, though they belong to the identical task. Based at the useful resource necessities, responsibilities are scheduled both withinside the identical or throughout distinctive servers. Usually, the company statistics the useful resource usage degrees of each scheduled challenge and maintains the person profiles.

The workload is the amount of work carried out with the help of employing a pc in every duration requested for many applications. The arrangement of this realities makes reasonable to outline the application's behavior and hone forecast techniques to discover out fate behaviors and estimate framework requests. In this way, the behaviors of workloads on the Cloud handling environment are emphatically connected with the CPU centers in comparison to Smash capacity of the machines on the server level. Hence, the mission asset utilization is as a rule communicated as multi-dimensional representation [6] including mission period in seconds, CPU utilization in centers, and memory utilization in gigabytes. It is commonly seen that most extreme of the designated CPU and memory asset are left unutilized all through mission execution. So, there may be a have to be explore the workload to diminish the utilization of the asset and computing cost.

The rest of this paper is organized as follows, in Section 2 explores various author's work on the workload prediction. The results are presented in Section 4 and finally the conclusions and future scope of the work are discussed in section 5.

2. Related work

Sergio Pachco-Sanchez et al [7] utilized Markovian Entry Forms (Outline) and related MAP/MAP/1 lining show as a device for execution of servers in cloud. By comparing with follow driven reenactment it was watched that Outline parameterization from HTTP log record leads to off-base expectation. They appeared that guess of lining behavior of follows can be way better accomplished by utilizing Most extreme Probability (ML) estimation.\

Kee kim et. al[8] compared the different forecast strategies to discover the leading suited forecast strategy for workload forecast. They compared workload forecast beneath real-world cloud arrangements. They assessed the combination of prescient – responsive (PR), Reactive-Predictive (RP) and Prescient –Predictive (PP) approaches. They concluded that no single strategy is all around best and recommended that prescient scaling –ve and +ve prescient scaling-out gives the leading comes about in term of fetched efficiency and lowest work due date miss rate. Rather than utilizing past information to anticipate end of the workload, the information about the workload of a pool of errands can be utilized. [9] proposed a strategy whereby the workload of existing errand are assembled in to numerous clusters, at that point neural arrange is utilized to memorize the characteristics of each cluster. At that point prepared neural organize is utilized to anticipate long haul workload as before long as the modern errand shows up.

Utilize of machine learning in workload forecast has moved forward the forecast capability. Different proactive provisioning strategies are utilized in cloud environment and their execution changes with the sort of workload.[10] compared the five major machine learning calculations in foreseeing the workload (CPU utilization).The execution of K-Nearest Neighbors (KNN), Straight Relapse (LR), Neural Arrange (NN) , Bolster Vector Machine (SVM) and Irregular Timberland (RF) were assessed .The execution changes with the workload sort and preparing. SVR gives the way better generally execution but at the taken a toll of higher preparing times. Time arrangement models can be utilized for stack predication. [11] Connected Autoregressive conditional Score to anticipate long run workload. The expectation demonstrate can be straight, nonlinear and crossover based upon the score characteristic of workload.

Padma D. Adane and O. G. Kakde [12] have done a comparative consider of Proactive provisioning approaches and Responsive provisioning approaches, they concluded that an in general made strides reaction time as the provisioning choices are taken some time recently the real require of assets emerge. The effectiveness of such proactive provisioning strategies is subordinate on the utilize of a prescient show that anticipates the asset prerequisites. In this paper we have assessed the execution of five prevalent Machine Learning Calculations in foreseeing the CPU utilization of different server logs taken from the Parallel Workload Document. The measurements utilized for assessment are MAE- Cruel Outright Mistake and RMSE- Root Cruel Squared Mistake

3. Methodology

3.1 Dataset Description

to get it and assess both the workload expectation and application situations, one must get to organize estimations from cloud systems. One must get it how the activity designs of cloud application workloads shift in arrange to create forecasts around them, additionally ought to be mindful how cloud systems change to put applications on them. Moreover, to legitimately assess Cicada, one ought to test its forecast and arrangement on genuine applications beneath genuine arrange conditions.

Earlier thinks about on datacenter systems have distinguished worldly and spatial inconstancy. Benson et al. [13] analyzed link-level SNMP logs from nineteen datacenters, although their applications may be comparative to those of cloud inhabitants. Benson et al. [14] assembled SNMP insights for ten datacenters and bundle follows from several switches in four datacenters. They portray a few of these as cloud information centers, but it is vague whether they are really IaaS networks.

Two datasets of web applications amassed setup. Datasets are ClarkNet [15], NASA [16]. The ClarkNet weblog was taken from a web server of Metro Baltimore–Washington, DC locale. The HTTP proxy logs were taken for two weeks from the web server. The dataset having 3,328,587 requests were observed in 2 weeks length. The

second dataset is NASA Kennedy Space Center and the server located in the Florida. This dataset having two months web logs totally 3,461,612 requests were observed

3.2 Description of Prediction Algorithms

The K-Nearest Neighbors (KNN) is one of the best of among ML algorithms and can be utilized for performance metrics. The K-most value plays important role to forecast comparative occasions when utilized for computation, in this algorithm [17]. It makes forecasts utilizing the preparing information set specifically. To decide the k comparable occurrences to an unused input its employments a separate degree.

The Linear Regression (LR) is often most fundamental strategy utilized in statistical analysis where all the qualities included within the expectation are numeric [18]. The yield to be anticipated is communicated as a straight with properties with foreordained weights. These are finds from the preparing information. For information logs with profoundly connected qualities, this algorithm performs with diminished precision [13].

The Support Vector Machine (SVM) is utilized for regression and too alluded to as Support Vector Regression (SVR). SVR tries to play down the mistake by finding a line of best fit [17]. It considers information occurrences closest to the least fetched line. Such occasions are known as Bolster Vectors. To oblige bended lines or polygon regions, it scales the information into higher measurements for forecasts. This will be accomplished by attempting out distinctive parts. SVM has the advantage of diminishing the issues of over-fitting or neighborhood minima [19].

The Random Forest (RF) is characterized in [17] is a generic guideline of classifier combination that employments L tree organized base classifiers $\{h(X, \Theta_n), N=1,2,3,\dots,L\}$, where X indicates the input information and $\{\Theta_n\}$ could be a family of indistinguishable and subordinate distributed random vectors. Each Decision Tree is made by randomly selecting the information from the accessible information. Random Forest can handle lost values and twofold information and consequently is reasonable for tall dimensional information modeling. It is effective, non-parametric and gives tall prediction accuracy [18].

The ARMA procedures are particularly reference estimators within the forecast of global radiation field. It could be a stochastic handle coupling autoregressive component (AR) to a moving average component (MA). This kind of show is commonly called ARMA (p, q) and is characterized with p and q parameters.

ARIMA stands for Auto Regressive Integrated Moving Normal. There are regular and Non-seasonal ARIMA models that can be utilized for estimating. Non-Seasonal ARIMA show, this strategy has three factors' Periods to slack (P) makes a difference alter the line that's being fitted to figure the arrangement, In an ARIMA demonstrate a time arrangement gets changed into stationary one utilizing differencing (D). D refers to the number of differencing changes required by the time arrangement to urge stationary. Q could be a variable that indicates the slack of the mistake component, where blunder component may be a portion of the time arrangement not clarified by drift or seasonality.

Regular ARIMA (SARIMA) models, $ARIMA(p,d,q) (P, D, Q)_m$ where p is the number of autoregressive, d is degree of differencing, q is the number of moving normal terms, m is alludes to the number of periods in each season, (P, D, Q) is speaks to the (p,d,q) for the regular portion of the time arrangement Regular differencing takes under consideration the seasons and contrasts the current value and it's esteem within the past season.

3.3 Experimental Setup

These experiments were conducted with 3.2 GHz speed and 16 GB RAM and the WEKA 3.9. The http proxy considered based on the format and created dataset having each row with eighteen attributes. The complete server information proxy has been changed over into tool native. Each information log has been cleaned and by sampling both evenly and after that vertically utilizing different channels accessible in Weka with preprocessing technique. Vertical inspecting strategy produces inputs which come near to real-world utilization of machine learning calculations.

In Weka, this can be known as highlight choice where each subset of qualities is assessed with the target machine learning calculation. The subset of qualities with ideal execution, multidimensionality and space information predisposition are chosen for expectation purposes. This techb decreases the number of traits considered for this assessment to five comprising of hold up time, run time, number of designated processors, normal CPU time utilized and utilized memory. For flat examining the cross-validation procedure has been

connected for evaluating the precision of the prediction models. Cross-validation is one of the foremost common mistake estimation procedures where each perception within the test dataset of measure n is progressively taken out and the remaining $n-1$ perceptions of the set are utilized to prepare the expectation show to gauge the anticipated asset usage.

The objective of this work was to assess the accuracy of the chosen machine learning procedures in anticipating the ask logs. The measurements utilized for assessment are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The Mean Absolute Error for the prediction is characterized as

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (1)$$

A littler RMSE esteem shows a more effective prediction conspire. The MAE perceptions made for each of the server logs for all the five machine learning calculations.

$$RSMA = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (2)$$

The Mean Squared Error (MSE) of an estimator measures the normal of the squares of the blunders that is, the normal squared distinction between the assessed values and the actual value. MSE could be a chance work, comparing to the anticipated esteem of the squared error loss.

$$MSE = \frac{\sum_{i=1}^n |(\hat{y}_i - y_i)|}{n} \quad (3)$$

The mean absolute percentage error (MAPE) may be a measurable degree of how accurate a forecast system is. It measures this exactness as a rate and can be calculated as the normal outright percent blunder for each time short genuine values separated by genuine values.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{(\hat{y}_i - y_i)^2}{y_i} \right| * 100 \quad (4)$$

4. Results

Table 1: Experiment results of NASA

Models	MAE	MSE	RMSE	MAPE
KNN	72.05	9542.26	95.32	20.63
LR	53.23	5764.74	75.63	15.23
SVM	68.62	10365.56	98.36	19.56
ARMA	88.96	14256.45	120.26	25.75
ARIMA	60.02	7256.75	85.63	16.65

Table 1 of NASA arrangement LR demonstrate gives the leading result from existing models. Consequently, the ARIMA show is essentially performing superior as compared to other models. The classification approach makes a difference to choose the suitable show with diverse workload design. SVM execution isn't up to the check for ClarkNet arrangement, and ARMA gives the most exceedingly bad execution within the case

Table 2: Experiment results of ClarkNet

Models	MAE	MSE	RMSE	MAPE
KNN	210.45	70265.12	250.81	12.45
LR	265.23	79638.23	295.14	15.24
SVM	250.63	95325.48	320.15	18.26
ARMA	220.3	86257.26	250.45	15.69
ARIMA	168.26	58234.18	235.72	11.52

Table 2 of ClarkNet arrangement appears the KNN demonstrate gives the leading comes about in existing models. Thus, the ARIMA demonstrate is altogether performing way better as compared to other models. The classification approach makes a difference to choose the fitting demonstrate with diverse workload design. ARMA

execution isn't up to the check for NASA arrangement, and SVM gives the most exceedingly bad execution within the case.

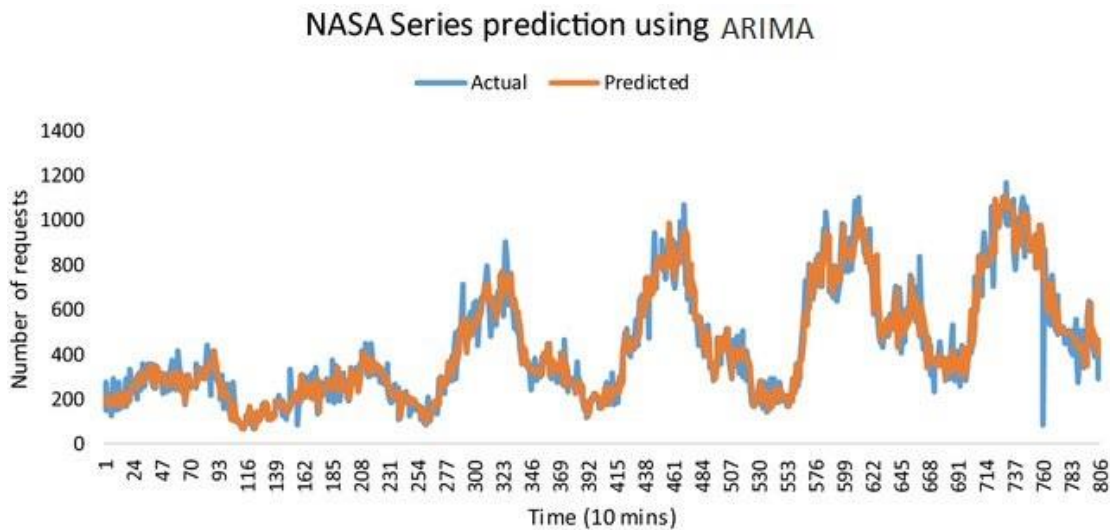


Figure 1: NASA series prediction using ARIMA

Figure 1 shows the prediction accuracy of ARIMA prediction model for NASA incoming workload. There is no significant difference for both actual and prediction.

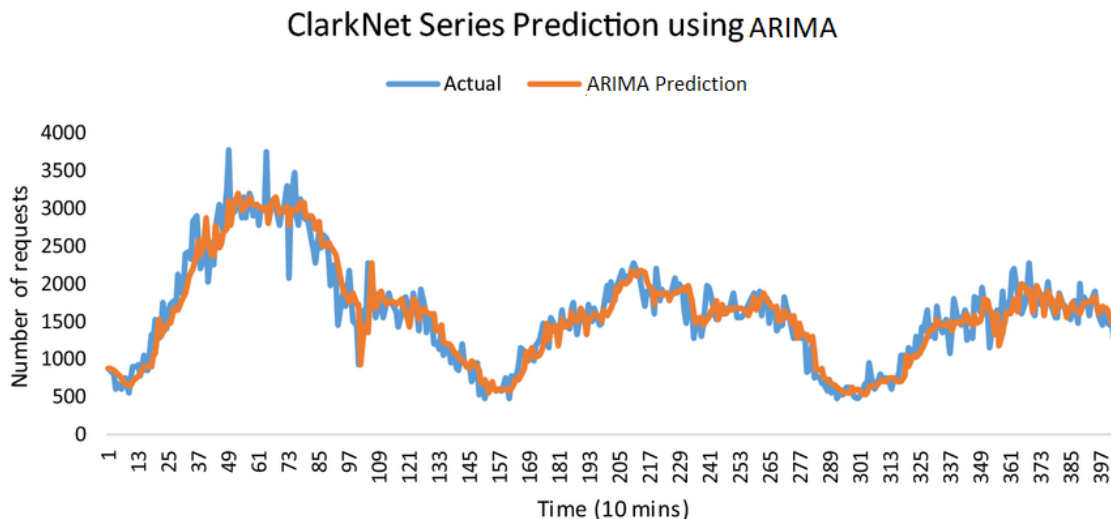


Figure 2: ClarkNet series prediction using ARIMA

Figure 2 shows the prediction accuracy of ARIMA prediction model for ClarkNet incoming workload. There is no significant difference for both actual and prediction.

5. Conclusions and Future Scope of Work

In cloud computing user pay only for the number of services used. Many models available for workload prediction of cloud environment till date are analytical or mathematical. Work can be done in future for automation of workload prediction of different cloud services. This paper compares different machine learning algorithms to predict the workload for future forecasting. For that ClarkNet and NASA datasets are used. The experimental results illustrate that LR and ARIMA model shows significant improvement for NASA and KNN and ARIMA shows significant improvement for ClarkNet. The QoS metrics have significant improvement; those are MAE, MSE, RMSE, MAPE. For both datasets the quality of service of web applications in a cloud environment and forecasting using ARIMA. This work can further be extended by applying other statistical methods besides this one can consider the usage machine learning techniques.

References

1. Gartner, Analyst Examine Top Industry Trends at Gartner Symposium/ITxpo, Orlando (2015).
2. Peter Mell, Timothy Grance: NIST Special Publication 800-145, "The NIST Definition of Cloud Computing", September 2011. <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>
3. Michael J. Kavis, Architecting the Cloud: Design Decision for Cloud Computing Service Models (SAAS, PAAS and IAAS), Wiley India Private Limited; 2014 edition, ISBN-10: 8126550333, ISBN-13: 978-8126550333.
4. David Rath August (2008). Cloud Computing: public-Sector Opportunities Emerge. Available at: <http://www.govtech.com/gt/articles/387269>.
5. P. A. Dinda and D. R. O'Hallaron, "Host load prediction using linear models," Cluster Computing, vol. 3, no. 4, pp. 265–280, 2000.
6. J. Panneerselvam, L. Liu, N. Antonopoulos, and Y. Bo, "Workload analysis for the scope of user demand prediction model evaluations in cloud environments," in Proceedings of the 7th IEEE/ACM International Conference on Utility and Cloud Computing (UCC '14), pp. 883–889, December 2014.
7. Sergio Pacheco-Sanchez, Giuliano Casale, Bryan Scoteny, Gerard Parr and Stephen Dawson. Markovian Workload Characterization for QoS Prediction in the Cloud. In 2011 IEEE 4th International Conference on Cloud Computing .978-0-7695-4460-1/11, 2011
8. Kee Kim, Wei Wang, Yanjun Qi and Marty Humphrey. Empirical Evaluation of Workload Forecasting Techniques for Predictive Cloud Resources Scaling. In 2016 IEEE 9th International Conference on Cloud Computing.
9. Yongila Yu, Vashu Jindal, I-Ling Yen, Farokh Bastani. Integrating Clustering and Learning for Improved Workload Prediction in the Cloud in 2016 IEEE 9th International Conference on Cloud Computing .215-6190/16, 2016 IEEE.
10. Padma D. Adane, O.G. Kakde. Predicting Resource utilization for Cloud Workloads Using Machine Learning Techniques. Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018). 978-1-5386-1974-2.
11. Abiola Adgboyega. Time-Series Models for Cloud Workload Prediction: A Comparison. In IFIP/IEEE International Symposium on Integrated Network Management (IM 2017). 978-3-901882-89-0.
12. Padma D. Adane and O. G. Kakde, Predicting Resource Utilization for Cloud Workloads Using Machine Learning Techniques, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore, ISBN:978-1-5386-1974-2.
13. Theophilus Benson, Ashok Anand, Aditya Akella, and Ming Zhang. Understanding Data Center Traffic Characteristics. In WREN, 2009.
14. Theophilus Benson, Aditya Akella, and David A. Maltz. Network Traffic Characteristics of Data Centers in the Wild. In IMC, 2010.
15. Balbach, S.: ClarkNet web server logs. <http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html> (2018). Accessed 25 Feb
16. Dumoulin, J.: NASA web server logs. <http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html> (2018). Accessed 25 Feb
17. Jason Brownlee, "Machine Learning Mastery with Weka," Ebook. Edition: v.1.4.
18. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.
19. N. Sapankevych and R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey," Computational Intelligence Magazine, IEEE, vol.4, no.2, pp.24-38, May 2009.