Optimizing Cloud Costs Using Predictive Analytics and Machine Learning Algorithms

Aina Sofea binti Amier Hamzah1, a) and Zarina Che Embi1, b)

1Faculty of Computing & Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia

b)Corresponding author: zarina.embi@mmu.edu.my

a) aina03sofea@gmail.com

**Abstract.** While cloud computing provides flexible and economical services, the unpredicted need for resources often leads to over or under provisioning which makes the service less cost effective. This research seeks to fill the gap of forecasting cloud computing costs accurately by analyzing the performance of regression models in comparison to time series forecasting models. This study compares regression models of decision tree, random forest, XGBoost, and neural network with time series models of RNN, LSTM, GRU, using a real world dataset of CPU, memory, network usage figures and cost metrics. Model evaluation is done using MAE, RMSE, NMAE, NRMSE, and R². The results demonstrate that time series models are more effective for capturing trends over time with metrics like CPU and memory utilization, while regression models outperform in analyzing static or sluggish data with observable multivariate relationships such as network data usage and cost metrics. These results show that model choice according to metric type improves accuracy, which leads to better strategies for managing cloud computing costs.

# Introduction

The transition from old technology to new technology has made many changes in improving the performance, quality, and cost-saving of a business, where one of the main contributions is cloud computing. Cloud computing is innovative in its approach. It uses scalable, flexible, and cost-effective solutions that permit access to software, platforms, and infrastructure on a pay-as-you-go basis [1]. This transforms the commercial approaches undertaken and fosters digital growth. On the other hand, cloud environments are subject to rapid changes that create complexities, especially in managing resources [2].

Unlike traditional systems in which resources are fixed and allocated for on-premises use, cloud computing environments must be adjusted proactively to serve the unpredictable level of demand from users [3]. This poses the risk of over-provisioning, and consequently, waste and higher spending, as well as under-provisioning, which causes a poor quality of service and a breach of service level agreements (SLA). Traditional approaches towards cost management, including statistical and probabilistic solutions, are simply inadequate for attending to the constantly fluctuating and multidimensional complexity of clouds [4][5].

Most cloud computing resources, such as CPU, memory, storage, and network bandwidth, need to be allocated precisely so that the operation remains cost-effective and meets the quality-of-service requirements. Cost over-provisioning is usually done to avoid SLA violations but leads to significant resource waste and increases in power usage, which is responsible for about 70% of the operational costs of data centers [6]. Besides, under-provisioning may negatively affect the quality of service, which can be a serious business concern. To illustrate, Amazon reportedly loses 1% of revenue for every 100 milliseconds (ms) increment in its response delay. In addition, Google suffers a 20% traffic decline for every additional 0.5-second delay in presenting search results.

In addition, using the cloud encompasses altercations in resource provisioning, computing, storage, and networking, which makes it complex and further requires precise resource management to ensure performance efficiency. More than 68% of organizations find it difficult to estimate cloud resource needs, exposing an opportunity gap that needs to be closed by novel approaches that respond to changing demands in real-time [7].

The combination of predictive analytics and machine learning techniques to solve these problems appears to be a valuable option. Analyzing past performance using predictive models enables an organization to optimize resource usage by cutting down wasted efforts and simultaneously reducing operational costs [8]. These methods involve a structured approach to achieve an optimal balance between costs and efficiency, and retain the agility and elasticity of the cloud solutions [9].

Even though multiple cost forecasting techniques are available for cloud environments, many of these solutions do not adequately address problems with ever-changing, multidimensional, or time-sensitive data. Traditional nonlinear forecasting and time series approaches struggle with univariate and multivariate interactions with nonlinearities, and regression techniques tend to ignore dependencies based on the time variable. This demonstrates a model construction gap, and this study attempts to find the appropriate model type to use based on the nature of the metric being forecasted. This research seeks to resolve this gap by analyzing the forecasting accuracy and predictive strength of regression estimation and time series approaches for multiple metrics on cloud resources. Accurate model identification can greatly enhance the efficiency of cloud resource management, aiding in cost and resource optimization, thereby allowing model selection processes for accurate and efficient cost optimization in cloud architecture.

# Literature review

Although there is wide coverage on research relating to the development of forecasting models for optimization of cloud costs, most researchers are focused on developing new algorithms or enhancing a single model’s effectiveness

(e.g., MSFS, HEADA, ProHPA, RAP). Most of these approaches attempt to isolate either time series models or machine learning regression models. However, there is a paucity of studies that benchmark the predictive efficiency of a broad range of model types on diverse cost-related cloud metrics.

This gap hinders cloud resource managers from selecting a model type suitable for a given prediction task. This study works on addressing this gap by conducting a benchmark comparison of regression and time series forecasting models on multiple cloud metrics (CPU, memory, network, and cost). Unlike previous studies that focus on constructing elaborate or hybrid algorithms, this paper aims to contribute the framework for model selection based on metric feature characterization ­­— temporal dependence versus multivariate dependence.

## Predictive Analytics

In the context of cloud environments, predictive analytics employs statistical and machine learning techniques to calculate resource allocations. These methods facilitate effective resource distribution, improving both operational efficiency and mitigating waste when scaling up (over-provisioning) and scaling down (under-provisioning), which can compromise performance, breach SLAs, or service level agreements [10][11].

Resource utilization forecasting has been done with classical time series models, including Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA). These models are useful in capturing short-term linear trends; however, they are extremely unscalable in complex cloud environments due to high-dimensional data’s non-linear dependencies [12].

The challenges can be overcome by utilizing deep learning based time series models such as RNNs, LSTMs, and GRUs. These architectures are capable of maintaining temporal dependencies and modeling more dynamic patterns within sequential data. Sequential data can be fed into these structures as input, which results in more accurately capturing the essence of temporal relationships contained in the nuances of the data. RNNs in essence are prone to suffering from gradient vanishing and exploding problems, thus hampering their learning capabilities regarding long-term dependencies [13]. LSTM and GRU models address this problem with the addition of memory cells and gating mechanisms [14][15][16]. LSTM models are more expressive of underlying patterns but are computationally expensive, while GRUs are less demanding in training time while maintaining performance. GRUs are thus better suited for dynamic forecasting of workload on clouds.

More advanced architectures have been attempted by prior work. For example, Xu et al. (2022) proposed supervised deep neural network (esDNN), which improves learning based on GRUs (and also weakening gradients among other things) by fixing gradient problems and enhancing gating procedures. In the same manner, L-PAW with TSA-encoding methods integrates Top-Sparse Auto-Encoders (TSA) into GRU for accurate real-time resource forecasting without excess computation during resource forecasting [13].

Although these advancements focus on accuracy, they widen the gap in multi-model comparative analysis since in forecasting, different models are usually studied in complete isolation—still, integrating results lacks robust foundational research. The objective of this work is to determine in which situations forecasting models such as RNN, GRU or LSTM outshine regression models and vice versa when using the same metrics and dataset for comparative analysis.

## Machine Learning

Due to the multidimensional nature of resource consumption data, machine learning is instrumental in modeling complex and nonlinear relationships which makes it pivotal for cloud cost optimization. Historical data-driven models such as Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and basic Neural Networks (NN) have learned to perform resource demand forecasting with keen accuracy [5][17].

These models work exceptionally well with multivariate inputs. In fact, they are ideal for estimating cloud workloads that do not strictly depend on time intervals. For example, Goodarzy et al. (2020) applied linear regression for CPU usage forecasting to turn virtual machines (VMs) on and off dynamically, which minimized the number of active VMs in use and improved resource efficiency [18]. Similary, random forests have been employed for workload prediction since the technique aggregates outputs from numerous decision trees improving accuracy and the robustness of the prediction [19].

The introduction of transformer-based time-series decomposition models [20] or GRU-BiLSTM [21] hybrids is without a doubt impressive. Nevertheless, their application is often limited to narrow, use-case scenarios guided by temporal patterns, which impedes their general utility across various cloud metrics. On the other hand, the regression models stand out as offering easy-to-interpret, low-cost, and readily available models suitable for decision-making in cost-sensitive environments. This study highlights these regression techniques due to their undeniable merits in practical scalability, resource-agnostic flexibility, ease of integration, and multi-facet adaptability, thus positioning them within frameworks aimed at optimizing costs in the cloud environment.

This study complements these specialized solutions by assessing general-purpose models along multiple metrics to ascertain which models serve best to each use case. Even though the models discussed above are targeted towards achieving a specific level of performance, there is a lack practical instruction for cloud administrators and researchers on what type of model corresponds to what metric (e.g., temporal vs. static data). This study aims to directly fill this gap.

In conclusion, the literature showcases a robust gap alongside an imbalance of focused attention devoted to the construction and refinement of highly tailored resource forecasting models for cloud resource management. That being said, there is insufficient cohesive benchmarking assessing regression and time series forecasting models against multiple resource metrics. This study aims to address this gap by determinedly studying which model types yield the best performance based on the attributes of the target metric. The results of this study are aimed at aiding practitioners in determining the most appropriate forecasting method for economically driven and strategically intelligent cloud resource distribution.

# MEthodology

The workflow of this study is shown in Figure 1. This framework integrates concepts from predictive analytics and machine learning to model usage trends, identify cost inefficiencies, and evaluate potential improvements. By applying advanced techniques such as time series forecasting and machine learning algorithms, the framework aims to enhance resource allocation and reduce costs while maintaining performance.

This study adopts a systematic approach which combines predictive analytics with machine learning to assess the performance of regression and time series forecasting models with respect to predicting cloud resource consumption and pertaining expenses [21].

The dataset, Cloud Cost Billing, is available on the Hugging Face platform [22] and contains 15 attributes with 124,275 instances. Some of the vital attributes are the cost data, along with CPU utilization, memory usage, and network throughput. Some data refinement techniques were performed, including the creation of new features through derived variables, so that model accuracy could be improved by eliminating some data inconsistencies.

Exploratory Data Analysis (EDA) actively guided the discovery of patterns, trends, and anomalies within the dataset. As for data preprocessing, techniques including normalization, log transformation, and missing value imputation were applied to ready the dataset for modeling.

Two sets of models were trained and assessed:

* Regression-based models: Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Neural Network (NN).
* Time series forecasting models: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).

Five evaluation metrics were applied: MAE, RMSE, NMAE, NRMSE, and R². These and others offer a wide-ranging perspective on the accuracy of the model, the error distribution and explanatory power over the given dataset, independent of scale or unit.

The objective was to evaluate accuracy and efficiency of the prediction for each resource type using regression and time-series forecasting in comparison to each other. The focus was not designing a hybrid model, but rather analyzing which approach is more appropriate based on the given metric’s features.

A diagram of a data analysis process

AI-generated content may be incorrect.

**FIGURE 1.** Project workflow

# Results

Results of predicting cloud resource usage using regression models and time series forecasting models are shown in Table 1. Table 1 shows the complete prediction performance for all models and includes various cloud resource metrics. Regression models and time series models have different strengths depending on the metric used.

For the case of CPU and Memory Utilization, which are temporally patterned metrics, the forecasting time series models, specifically GRU and LSTM, yielded the best results for accuracy, MAE and RMSE, and R² scores (0.9999). These models have captured dependencies of time-based factors very competitively. Therefore, they are applicable to resource metrics with dynamic resource nature.

For the remaining metrics like Network Data and Rounded Cost, time series models were outperformed by regression models. Quite a few of these models demonstrated less time-dependency such as Random Forest and Decision Tree which focus on learning complex relationships among multiple variables rather than on a sequence-based input.

Cost Per Quantity had a slightly better performance with LSTM which suggests to us that some cost-type metrics might be helped by time series models depending on the degree of their fluctuation.

The best-performing model for every metric is marked in Table 2. The emphasized result confirms once again that there is no model type dominant across all categories. Model selection should be based on the characteristics of the data and the goal of prediction.

These findings support the usefulness of a comparative approach to model selection: time series models perform best in dynamic, sequential contexts, while regression models are effective with static or structurally patterned data.

**TABLE 1.** Results evaluation across all metrics and models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predicted Metrics** | **Type of Model** | **Models** | **MAE** | **RMSE** | **NMAE** | **NRMSE** | **R²** |
| CPU Utilization | Regression Model | DT | 0.2428 | 0.3115 | 0.0024 | 0.0031 | 0.9998 |
| RF | 0.0742 | 0.0981 | 0.0007 | 0.0009 | 0.9999 |
| XGB | 0.5399 | 0.6977 | 0.0054 | 0.0070 | 0.9994 |
| NN | 0.0064 | 0.0102 | 6.44e-05 | 0.0001 | 0.9999 |
|  | Time Series Forecasting Model | RNN  LSTM  GRU | 0.0107  0.0012  0.0012 | 0.0111  0.0015  0.0015 | 0.0213  0.0025  0.0025 | 0.0221  0.0031  0.0030 | 0.99190.9999  0.9999 |
| Memory Utilization | Regression Model | DT | 0.2395 | 0.3081 | 0.0024 | 0.0031 | 0.9998 |
| RF | 0.0749 | 0.0985 | 0.0007 | 0.0010 | 0.9999 |
| XGB | 0.5448 | 0.7049 | 0.0054 | 0.0070 | 0.9994 |
| NN | 0.0066 | 0.0099 | 6.60e-05 | 0.0001 | 0.9999 |
|  | Time Series Forecasting Model | RNN  LSTM  GRU | 0.0048  0.0021  0.0009 | 0.0052  0.0025  0.0011 | 0.0095  0.0041  0.0017 | 0.0103  0.0049  0.0021 | 0.9983  0.9996  0.9999 |
| Network Data | Regression Model | DT | 0.0008 | 0.0013 | 8.91e-06 | 1.33e-05 | 0.9999 |
| RF | 0.0005 | 0.0007 | 5.71e-06 | 7.79e-06 | 0.9999 |
| XGB | 0.0889 | 0.1036 | 0.00089 | 0.0010 | 0.9999 |
| NN | 0.0096 | 0.0114 | 9.63e-05 | 0.0001 | 0.9999 |
| Time Series Forecasting Model | RNN  LSTM  GRU | 0.0014  0.0025  0.0008 | 0.0019  0.0027  0.0010 | 0.0028  0.0054  0.0020 | 0.0037  0.0054  0.0020 | 0.9998  0.9995  0.9999 |
| Cost Per Quantity | Regression Model | DT | 0.0370 | 0.0552 | 0.0037 | 0.0055 | 0.9996 |
| RF | 0.0124 | 0.0236 | 0.0012 | 0.0023 | 0.9999 |
| XGB | 0.0822 | 0.1131 | 0.0083 | 0.0114 | 0.9984 |
| NN | 0.1032 | 0.1564 | 0.0104 | 0.0158 | 0.9969 |
| Time Series Forecasting Model | RNN  LSTM  GRU | 0.0030  0.0018  0.0018 | 0.0036  0.0021  0.0022 | 0.0059  0.0035  0.0036 | 0.0073  0.0042  0.0044 | 0.9991  0.9997  0.9997 |
| Rounded Cost | Regression Model | DT | 2.3e-05 | 0.0007 | 2.43e-06 | 7.49e-05 | 0.9999 |
| RF | 4.1e-05 | 0.0008 | 4.27e-06 | 8.96e-05 | 0.9999 |
| XGB | 0.0066 | 0.0181 | 0.0006 | 0.0019 | 0.9997 |
| NN | 0.0054 | 0.0071 | 0.0005 | 0.0007 | 0.9999 |
| Time Series Forecasting Model | RNN  LSTM  GRU | 0.0025  0.0012  0.0024 | 0.0029  0.0013  0.0027 | 0.0148  0.0074  0.0036 | 0.0171  0.0080  0.0044 | 0.9991  0.9995  0.9999 |

**TABLE 2.** Summary of the best performing models for each metrics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Predicted Metric** | **Best Model Type** | **Best Model** | **MAE** | **RMSE** | **NMAE** | **NRMSE** | **R²** |
| CPU Utilization | Time Series Forecasting | GRU | 0.0012 | 0.0015 | 0.0025 | 0.0030 | 0.9999 |
| Memory Utilization | Time Series Forecasting | GRU | 0.0009 | 0.0011 | 0.0017 | 0.0021 | 0.9999 |
| Network Data | Regression | RF | 0.0005 | 0.0007 | 5.71e-06 | 7.79e-06 | 0.9999 |
| Cost Per Quantity | Time Series Forecasting | LSTM | 0.0018 | 0.0021 | 0.0035 | 0.0042 | 0.9997 |
| Rounded Cost | Regression | RF | 4.1e-05 | 0.0008 | 4.27e-06 | 8.96e-05 | 0.9999 |

# CONCLUSION

This study performed a comparative evaluation of regression-based forecasting and time series forecasting cloud resource usage forecasting models with the intent to optimize costs. Different cloud metric types utilize differently structured models, which proved in the evaluation with MAE, RMSE, NMAE, NRMSE, and R².

For example, CPU and memory usage were best predicted by time series models GRU and LSTM. Less dynamic metrics such as network data and cost also utilized more efficient forecasting through regression models like Random Forest and Decision Tree.

By tailoring model selection to the characteristic of the metric, the study demonstration that predictive accuracy and cost efficiency within cloud resource optimization could be advanced greatly. These insights greatly benefit cloud service providers and organizations striving for intelligent, data-driven expense management frameworks.

While this research successfully applies regression-based models for cloud cost forecasting and resource leveraging, several areas for additional work still exist. A primary focus is to improve responsiveness over time regarding the use of regression models by applying more sophisticated feature design processes like automated lag selection and rolling statistical indicators. These methods will enhance capturing attention to changing workload patterns.

Furthermore, future work can explore modeling hybrids that merge the regression model's interpretative generalizability with the strength of deep learning sequence models LSTM or Transformer networks. This would permit responsive adaptive forecasting where workload patterns shift dynamically and decisions need instant, time-sensitive triggers.

Finally, embedding the future work models into a real cloud orchestration system and evaluating their impact on actual cost, performance, and energy consumption metrics would test scalability and practical deployment usefulness.

# Acknowledgments

The authors would like to thank Multimedia University for the support given to this work.

# References

1. A.K. C, S. B, and N. R, “Resource Utilization Prediction in Cloud Computing using Hybrid Model,” IJACSA **12**(4), (2021).
2. S. A. Zargar, “Introduction to sequence learning models: RNN, LSTM, GRU,” (2021). https://doi.org/10.13140/RG.2.2.36370.99522
3. J. Samual, T. Alkhdour, A. K. Wiesi, M. Amin Almaiah, A. Ali, R. Al-Ali, T. H. H. Aldhyani, and R. Shehab, “Optimizing cloud resource utilization through machine learning forecasting,” *J. Theor. Appl. Inf. Technol.* **15**(17), (2024). Available at: https://www.researchgate.net/publication/385750605
4. M. Darwich, T. Alghamdi, and M. Bayoumi, “Deep Learning Approach for Cost and Storage Optimization of Video Streaming in Cloud Environments,” in *2023 IEEE 8th International Conference on Smart Cloud (SmartCloud)*, (IEEE, Tokyo, Japan, 2023), pp. 80–85.
5. P. Gaba, Himanshu, Preetitanya, and Y. Gupta, “Unlocking Efficiency - Multidimensional Cost Optimization Strategies for Cloud Infrastructure in Small and Medium-Sized Organizations,” in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, (IEEE, Pudukkottai, India, 2023), pp. 463–470.
6. S. Goodarzy, M. Nazari, R. Han, E. Keller, and E. Rozner, “Resource Management in Cloud Computing Using Machine Learning: A Survey,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (IEEE, Miami, FL, USA, 2020), pp. 811–816.
7. A. Trivedi, “Available online,” *Int. J. Res. Radic. Multidiscip. Fields* (**IJRRMF**) **3**(2), (2024).
8. M.F. Manzoor, A. Abid, M.S. Farooq, N. A. Azam, and U. Farooq, “Resource Allocation Techniques in Cloud Computing: A Review and Future Directions,” ELEKTRON ELEKTROTECH **26**(6), 40–51 (2020).
9. A.Q. Khan, M. Matskin, R. Prodan, C. Bussler, D. Roman, and A. Soylu, “Cloud storage cost: a taxonomy and survey,” World Wide Web **27**(4), 36 (2024).
10. M. Ade, “Optimizing cloud resource provisioning with machine learning,” (2024). Available at: <https://www.researchgate.net/publication/384766637>
11. V. N. Tsakalidou, P. Mitsou, and G. A. Papakostas, “Machine learning for cloud resources management—An overview,” (n.d.).
12. M. Smendowski, and P. Nawrocki, “Optimizing multi-time series forecasting for enhanced cloud resource utilization based on machine learning,” Knowledge-Based Systems **304**, 112489 (2024).
13. Z. Chen, J. Hu, G. Min, A. Y. Zomaya, and T. El-Ghazawi, “Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning,” *IEEE Trans. Parallel Distrib. Syst.* **31**(4), 923–934 (2020). https://doi.org/10.1109/TPDS.2019.2953745
14. K. Bandara, C. Bergmeir, and H. Hewamalage, “LSTM-MSNet: Leveraging Forecasts on Sets of Related Time Series With Multiple Seasonal Patterns,” IEEE Trans. Neural Netw. Learning Syst. **32**(4), 1586–1599 (2021).
15. M. Xu, C. Song, H. Wu, S.S. Gill, K. Ye, and C. Xu, “esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments,” ACM Trans. Internet Technol. **22**(3), 1–24 (2022).
16. H. Yadav, and A. Thakkar, “NOA-LSTM: An efficient LSTM cell architecture for time series forecasting,” Expert Systems with Applications **238**, 122333 (2024).
17. N. Puso, T. Sigwele, and O.Z. Mustapha, “Machine Learning Centered Energy Optimization In Cloud Computing: A Review,” IJEEI **11**(3), 834–853 (2023).
18. S. Goodarzy, M. Nazari, R. Han, E. Keller, and E. Rozner, “Resource Management in Cloud Computing Using Machine Learning: A Survey,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (IEEE, Miami, FL, USA, 2020), pp. 811–816.
19. V. Khetani, Y. Gandhi, S. Bhattacharya, S. N. Ajani, and S. Limkar, “Cross domain analysis of ML and DL: Evaluating their impact in diverse domains,” *Int. J. Intell. Syst. Appl. Eng.* **2023**(7s), (n.d.). Available at: www.ijisae.org.
20. J. Zhu, W. Bai, J. Zhao, L. Zuo, T. Zhou, and K. Li, “Variational mode decomposition and sample entropy optimization based transformer framework for cloud resource load prediction,” Knowledge-Based Systems **280**, 111042 (2023).
21. N. Sairam (2024, September). sairamn/gcp-cloud-billing-cost · Datasets at Hugging Face. https://huggingface.co/datasets/sairamn/gcp-cloud-billing cost