International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Forecasting Cloud Spot Instances Using Machine Learning Models: A Review

Arti Parge[1], Prof. Sanmati Jain[2]

Abstract: Cloud services are worldwide used in different kinds of applications. Thereby on these servers are depends upon the use in application and the traffic on applications. In order to execute these services for the applications, the computational resources are required. These computational resources are termed as the Amazon spot instance. As the load and demand of these spot instances varies the prices of these resources are also varies. Additionally, the resources and their utilization in different applications are not predictable due to dependency upon other applications and the demands. In order to understand the price variation for spot instance prices the prediction algorithms are required. The data mining technique provides
us the ability to learn with the historical data trends and identify or recognize the similar trends of the data. This ability makes it essential for various new generation applications. This paper presents a comprehensive review on Data Driven Approaches for optimal prediction of cloud spot instances.

Keywords: Cloud Computing, Elastic Cloud, Spot Instances, Regression Analysis.

I. INTRODUCTION

Cloud computing has become highly pervasive technology in recent years [1]. A number of companies and organizations are now-a-days hosting their data and applications over cloud. Basically in order to run or execute applications over cloud expensive resources are required. Additionally the demand of resources is varying continuously, due to variations in load time or demand of computing resources [2]. Therefore the prices of the resources are also varying according to the time, demand and load on resources. Therefore the cost of resources cannot be determined directly and bidding process is used to buy the resources according to the needs.

In this context, for making effective bids for purchasing the resources in best prices the predictive methodologies can help [3]. Thus in the presented work the approach is proposed using the data mining technique that is used for predicting the price of computational resources. The data mining techniques are applied to computational algorithms on the historical price data [4].

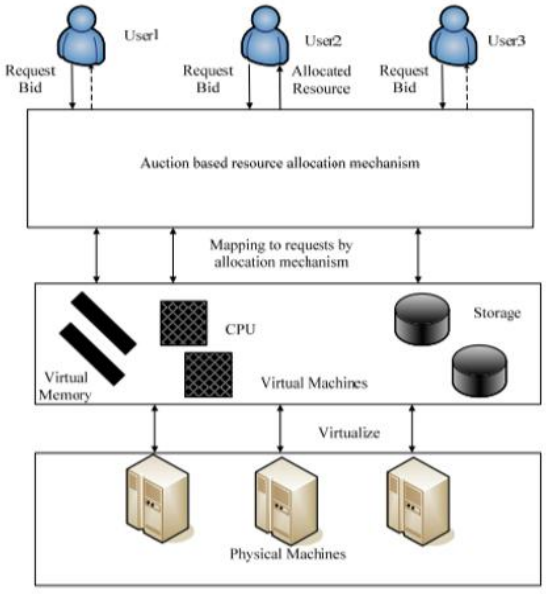

Figure.1 The Spot Instances Framework

For the prediction, classification, clustering data mining techniques are used. All these techniques are applied for finding the trends on data or relationship in data attributes. Using these algorithms trends of price varies is modeled using the data models, through regression analysis [6].

II. CLOUD SPOT INSTANCES

The concept of elastic cloud and spot instances is very crucial for cloud based platforms. The use of cloud computing has become increasingly popular due to its scalability, flexibility, and cost-effectiveness. Cloud service providers, such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure, offer various pricing models, including on-demand instances and spot instances [7]. Spot instances are a cost-saving option that allows users to bid for spare cloud computing capacity. This report focuses on estimating cloud spot instances and explores the key considerations and strategies involved [8]. Cloud spot instances are surplus computing resources made available by cloud service providers at significantly reduced prices. They are suitable for non-critical workloads, such as batch processing, data analysis, and testing environments. However, spot instances are subject to availability and may be interrupted if the bid price exceeds the current market price [9].

Factors Affecting Spot Instances:

The following factors effect spot instances the most [11]-[12]:

a. Supply and Demand: Spot instance prices fluctuate based on the supply of available instances and the demand from users. High demand can lead to increased prices, while low demand may result in lower prices.

b. Instance Type: Different instance types have varying spot instance prices based on their capabilities, specifications, and popularity.

c. Region and Availability Zone: Spot instance prices can vary across different regions and availability zones due to factors like data center location, infrastructure costs, and demand dynamics.

d. Time of Day: Spot instance prices may vary depending on the time of day, as certain periods experience higher demand, such as business hours or peak processing times [12].

Estimating Spot Instances:

To estimate spot instance costs effectively, the following strategies can be employed [13]:

1. **Historical Pricing Analysis:** Analyzing historical spot instance pricing data can provide insights into price trends, seasonal patterns, and peak demand periods. This analysis helps optimize bidding strategies and budget allocation.

2. **Real-Time Spot Price Monitoring:** Continuous monitoring of spot instance prices allows users to identify fluctuations and make informed decisions about bidding or launching instances during periods of low prices [14].

3. **Bidding Strategies:** Implementing intelligent bidding strategies is crucial to securing spot instances at optimal prices. Techniques such as setting bid prices below on-demand prices, utilizing automated bidding tools, and setting bid thresholds can help increase the chances of obtaining spot instances [15].

4. **Utilization Metrics:** Monitoring instance utilization metrics can help identify idle or underutilized instances, allowing users to take advantage of spot instances effectively [16].

III. PREVIOUS WORK

This section presents the noteworthy work in the domain: Nezamdoust et al. proposed that the cost fluctuates dynamically based on supply and demand. "Spot price" is a common term for this. To be able to use this instance, the user must create a suitable offer above the spot price. Accurate spot price prediction allows users to pre-prepare bid prices and run time to increase the reliability of the method. For this purpose, we consider Amazon EC2 as a testbed and use its spot price history to predict the future price by constructing a proposed modified gated recurrent unit (MGRU) model and providing a proposed dropout method. Compared with other sophisticated methods, test results show that the proposed method works superior and more accurately.

Lee et al. proposed that he price dataset, the recently introduced spot instance availability and interruption

ratio datasets can help users better utilize spot instances, but they are rarely used in reality. With a thorough analysis, we could uncover major hurdles when using the new datasets concerning the lack of historical information, query constraints, and limited query interfaces. To overcome them, we develop SpotLake, a spot instance data archive web service that provides historical information of various spot instance datasets. Novel heuristics to collect various datasets and a data serving architecture are presented. Through real-world spot instance availability experiments, we present the applicability of the proposed system. SpotLake is publicly available as a web service to speed up cloud system research to improve spot instance usage and availability while reducing cost.

Katayama et al. proposed that Amazon Web Services (AWS) provides its excess computing resources as spot instances. Although spot instances are cheaper than on-demand instances, they may be removed by AWS when demand increases for computing resources. Consumers need to judiciously select and prioritize those resources whose operations must avoid being interrupted. Although recently a spot placement score (SPS) is being provided as a measure of spot instance availability, few researches are using SPS. In this study, we propose an instance selection algorithm using it to select spot instances that reduce interruptions.

Zhou et al. proposed that recent update of spot pricing model on Amazon EC2, these work may become either inefficient or invalid. In this article, we present FarSpot which is an optimization framework for HPC applications in the latest cloud spot market with the goal of minimizing application cost while ensuring performance constraints. FarSpot provides accurate long-term price prediction for a wide range of spot instance types using ensemble-based learning method. It further incorporates a cost-aware deadline assignment algorithm to distribute application deadline to each task according to spot price changes. With the assigned subdeadline of each task, FarSpot dynamically migrates tasks among spot instances to reduce execution cost. Evaluation results using real HPC benchmark show that 1) the prediction error of FarSpot is very low (below 3%), 2) FarSpot reduced the monetary cost by 32% on average compared to state-of-the-art algorithms, and 3) FarSpot satisfies the user-specified deadline constraints at all time.

Portella et al. proposed that the price of virtual machine instances in the Amazon EC2 spot model is often much lower than in the on-demand counterpart. However, this price reduction comes with a decrease in the availability guarantees. Several mechanisms have been proposed to analyze the spot model in the last years, employing different strategies. To our knowledge, there is no work that accurately captures the trade-off between spot price and availability, for short term analysis, and does long term analysis for spot price tendencies, in favor of user decision making. In this work, we propose (a) a utility-based strategy, that balances cost and availability of spot instances and is targeted to short-term analysis, and (b) a LSTM (Long Short Term Memory) neural network framework for long term spot price tendency analysis.

III. EXISTING MODELS

The major challenge with design of health risk prediction systems are [17]:

1) The data is extremely complex and uncorrelated in nature.
2) The number of variables being large makes it extremely challenging to carry out regression analysis.
3) The outcomes are often individual dependent not exhibiting alignment to fixed patterns.

Mostly, evolutionary algorithms are used in the domain to design models for health risk prediction**.** Evolutionary algorithms try to mimic the human attributes of thinking which are:

1) Parallel data processing
2) Self-Organization
3) Learning from experiences

The major approaches employed in the domain of health risk prediction are:

Some of the commonly used techniques are discussed below:

1) Statistical Regression: These techniques are based on the time series approach based on the fitting problem that accurately fits the data set at hand. The approach generally uses the auto-regressive models and means statistical measures. They can be further classified as [18]:

a) Linear

b) Non-Linear

Mathematically:

Let the time series data set be expressed as:

$$Y = \{Y1, Y2 \dots \dots \dots \dots \dots \dots \dots Yt) \quad (1)$$

Here,

Y represents the data set

t represents the number of samples

Let the lags in the data be expressed as the consecutive differences.

The first lag is given by:

$$\Delta Y_1 = Y_{t-1} \quad (2)$$

Similarly, the j^{th} lag is given by:

$$\Delta Y_j = Y_{t-j} \quad (3)$$

2) Correlation based fitting of time series data: The correlation based approaches try to fit the data based on the correlation among the individual lags. Mathematically it can be given by [18]:

$$A_t = corr(Y_t, Y_{t-1}) \quad (4)$$

Here,

Corr represents the auto-correlation (which is also called the serial correlation)

Y_t is the t^{th} lagged value

Y_{t-1} is the $(t-1)^{st}$ lagged value

The mathematical expression for the correlation is given by

$$corr(Y_t, Y_{t-1}) = \frac{conv(Y_t, Y_{t-1})}{\sqrt{varY_t, varY_{t-1}}} \quad (5)$$

Here,

Conv represents convolution given by:

$$conv\{x(t), h(t)\} = \int_{t=1}^{\infty} x(\vartheta)h(t-\vartheta)d\vartheta \quad (6)$$

μ_t is the time dependent combination-coefficient

4) Artificial Neural Networks (ANN) and Deep Neural Networks (DNNs): In this approach, the time series data is fed to a neural network resembling the working of the human based brain architecture with a self-organizing memory technique [19].

The approach uses the ANN and works by training and testing the datasets required for the same. The general rule of the thumb is that 70% of the data is used for training and 30% is used for testing. The neural network can work on the fundamental properties or attributes of the human brain i.e. parallel structure and adaptive self-organizing learning ability. Mathematically, the neural network is governed by the following expression:

$$Y = f(\sum_{i=1}^{n} X_i . W_i + \theta_i) \quad (7)$$

Here,

X_i represents the parallel data streams

W_i represents the weights

θ represents the bias

f represents the activation function.

The second point is critically important owing to the fact that the data in time series problems such as sales forecasting may follow a highly non-correlative pattern and pattern recognition in such a data set can be difficult.

Mathematically:

$x = f(t)$

Here,

x is the function

t is the time variable.

The relation f is often difficult to find being highly random in nature.

The neural network tries to find the relation f given the data set (D) for a functional dependence of x(t).

The data is fed to the neural network as training data and then the neural network is tested on the grounds of future data prediction. The actual outputs (targets) are then compared with the predicted data (output) to find the errors in prediction. Such a training-testing rule is associated for neural network. Deep Neural Networks are the neural networks with multiple hidden layers and are generally used for training complex datasets.

IV. EVALUATION PARAMETERS

The performance metrics of the machine learning based classifier is generally done based on:

The parameters which can be used to evaluate the performance of the ANN design for time series models is given by:

1) Mean Absolute Error (MAE)
2) Mean Absolute Percentage Error (MAPE) and
3) Mean square error (MSE)

The above mentioned errors are mathematically expressed as:

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |V_t - \hat{V}_t| \quad (8)$$

Or

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |e_t| \quad (9)$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N} \frac{|V_t - \hat{V}_t|}{V_t} \quad (10)$$

The mean square error (MSE) is given by:

$$MSE = \frac{1}{N}\sum_{t=1}^{N} {e_t}^2 \quad (11)$$

Here,

N is the number of predicted samples
V is the predicted value
$\hat{V}_t$ is the actual value
e is the error value

Conclusion: Estimating cloud spot instances involves understanding various factors influencing pricing, implementing effective bidding strategies, and mitigating interruptions. By utilizing historical pricing analysis, real-time monitoring, and intelligent bidding approaches, users can optimize costs and maximize the benefits of spot instances while considering fault-tolerant architectures to minimize disruptions. Cloud spot instances offer an attractive option for cost-conscious organizations seeking cost savings without compromising workload performance and scalability. This paper presents a comprehensive review on estimating spot instances along with their salient featues.

References:

1. SS Nezamdoust, MA Pourmina, F Razzazim, " Optimal prediction of cloud spot instance price utilizing deep learning" The Journal of Supercomputing, Springer 2023, vol.79, pp.– 7647.
2. S. Lee, J. Hwang and K. Lee, "SpotLake: Diverse Spot Instance Dataset Archive Service," 2022 IEEE International Symposium on Workload Characterization (IISWC), Austin, TX, USA, 2022, pp. 242-255.
3. D. Katayama, K. Kasai and T. Koita, "Migration Destination Selection Algorithm for Spot Instances using SPS," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 6690-6692
4. A. C. Zhou, J. Lao, Z. Ke, Y. Wang and R. Mao, "FarSpot: Optimizing Monetary Cost for HPC Applications in the Cloud Spot Market," in IEEE Transactions on Parallel and Distributed Systems, 2021, vol. 33, no. 11, pp. 2955-2967
5. G. J. Portella, E. Nakano, G. N. Rodrigues, A. Boukerche and A. C. M. A. Melo, "A Novel Statistical and Neural Network Combined Approach for the Cloud Spot Market," in IEEE Transactions on Cloud Computing, 2023 vol. 11, no. 1, pp. 278-290.
6. 6. Liu D, Cai Z, Lu Y (2019) Spot price prediction based dynamic resource scheduling for web applications. In: 2019 Seventh International Conference on Advanced Cloud and Big Data (CBD). IEEE, pp 78–83 7.
7. Varshney P, Simmhan Y (2019) AutoBot: Resilient and cost-effective scheduling of a bag of tasks on spot VMs. IEEE Trans Parallel Distrib Syst 30(7):1512-1527
8. Sharma P, Lee S, Guo T, Irwin D, Shenoy P (2017) Managing risk in a derivative IaaS cloud. IEEE Trans Parallel Distrib Syst 29(8):1750-1765
9. Mishra AK, Yadav DK (2017) Analysis and prediction of Amazon EC2 spot instance prices. Int J Appl Eng Res 12(21):11205– 11212
10. Teylo L, Arantes L, Sens P, Drummond LM (2019) A bag-of-tasks scheduler tolerant to temporal failures in clouds. In: 2019 31st International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). IEEE, pp. 144–151

11. Khandelwal V, Chaturvedi AK, Gupta CP (2020) Amazon EC2 spot price prediction using regression random forests. IEEE Trans Cloud Comput 8(1):59–72
12. Khashei M, Bijari M (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput 11(2):2664-2675.
13. Liu Y, Wang Z, Zheng B (2019) Application of regularized GRU-LSTM model in stock price prediction. In: 2019 IEEE 5th International Conference on Computer and Communications (ICCC). IEEE, pp 1886-1890
14. Abbasimehr H, Shabani M, Yousefi M (2020) An optimized model using LSTM network for demand forecasting. Comput Ind Eng 143(106435):1-13.
15. DAI G, MA C, XU X (2019) Short-term traffic flow prediction method for urban road sections based on space-time analysis and GRU. IEEE Access 7(1):143025-143035
16. Song J, Tang S, Xiao J, Wu F, Zhang Z (2016) LSTM-in-LSTM for generating long descriptions of images. Comp Visual Media 2(4):379–388.
17. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929-1958.
18. Wang Z, Zhu R, Zheng M, Jia X, Wang R, Li T (2019) A regularized LSTM network for short-term traffic flow prediction, In: 2019 6th International Conference on Information Science and Control Engineering (ICISCE). IEEE, pp 100-105
19. Singh VK, Dutta K (2015) Dynamic price prediction for Amazon spot instances, In: 2015 48th Hawaii International Conference on System Sciences (HICSS). IEEE, pp 1513–1520

.