AWS EC2 Instance Spot Price Forecasting Using LSTM Networks

Yejur Kunwar, Jeffrey Lancon, David Stroud, Monnie McGee, Robert Slater

Southern Methodist University, 6425 Boaz Lane,
Dallas, Texas 75205
{ykunwar,jlancon,jdstroud,mmcgee,rslater}@smu.edu

Abstract. Cloud computing is a network of remote computing resources hosted on the Internet that allow users to utilize cloud resources on demand. As such, it represents a paradigm shift in the way businesses and industries think about digital infrastructure. With the shift from IT resources being a capital expenditure to a managed service, companies must rethink how they approach utilizing and optimizing these resources in order to maximize productivity and minimize costs. With proper resource management, cloud resources can be instrumental in reducing computing expenses.

Cloud resources are perishable commodities; therefore, cloud service providers have developed strategies to maximize utilization of their resources. One method cloud providers employ is offering unused/excess computing resources at substantially discounted rates compared to other pricing tiers, whose pricing fluctuates with supply and demand levels. This is often referred to as spot pricing.

This study investigates methods to reduce risk and increase predictability of pricing for businesses utilizing Amazon Web Services (AWS) elastic compute cloud (EC2) Spot instance pricing tier by accurately predicting spot instance pricing over a specified time-frame using long short-term memory (LSTM) neural networks and comparing the results against traditional time-series Auto Regressive Integrated Moving Average (ARIMA) modeling. The results show LSTM model Spot Instance price predictions have an average reduction in mean absolute percent error (MAPE) of approximately 95 percent when compared to the baseline ARIMA model.

1 Introduction

Cloud computing is state of the art way in the way businesses and the IT industry think about digital infrastructure. The days when businesses purchase and maintain private server farms and networks for their needs are being replaced with cloud computing providers. From multinational online retailers to mid-size industrial corporations to sole proprietor craft stores, all are utilizing or looking to utilize cloud computing services.

Cloud computing is on-demand computing resources that enable businesses to focus on their core competencies. Cloud computing is a model for enabling

ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]."

Cloud computing is a subscription model, where user's only pay for the resources they need when they need them and terminate them when the resource is no longer required. This improves users agility and flexibility when responding to increasing and/or highly volatile computing resource requirements. Resources can be quickly, often automatically, provisioned when demand is high and eliminated when the demand is less. This is in stark contrast to when capacity was governed by the number of servers and network capacity that a business owned. To provision for peak loading, businesses often over-invest in IT infrastructure capacity, increasing capital expenditures(CAPEX). Cloud computing delivery model converts capital expenditures (e.g. purchasing servers, routers, increased bandwidth) to an operational expenditure (OPEX) (e.g. hourly fees for usage, storage).

The switch from capital expenditure model to operational expenditure model for computing resources is having a profound impact on how businesses think about computing resources. Switching from computing resources being considered a cost of doing business overhead, to a measured service, where resources are allocated, metered, and optimized according to the tasks being conducted, allows businesses to properly allocate and optimize resource costs. Cloud computing resource utilization can be monitored, reported and controlled real-time. Companies whose business model depends on computing resources are increasingly adopting a cloud-first strategy within their organizations, opting to utilize shared cloud computing resources instead of hosting their own private computing resources. Reasons given for moving to cloud computing resources is the ability to scale workloads up and down as needed as well as costs savings over private networks [2].

A key cloud computing characteristic is rapid elasticity, provisioning and releasing resources commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time [1].

Cloud services are often segmented into three different types of offerings; Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS). Infrastructure as a Service (IaaS) is a service model that provides computing infrastructure; compute, storage, network capabilities, and basic software to clients [such as Amazon Web Services (AWS), Google Compute Engine (GCE),]. Platform as a Service (PaaS) is a service model that allows client to use fully functioning products and services to run existing applications [such as MySql, MongoDB, Windows Azure], or host custom designed applications. Software as a Service (SaaS) utilizes the cloud to deliver complete applications that are managed by 3rd party vendors and accessed by users via web interfaces [such as Salesforce, GoToMeeting, Google Apps,]. For the scope of this analysis, we will be focusing solely on Infrastructure as a Service (IaaS).

A recent public cloud service provider revenue forecasts showed that Infrastructure as a Service (IaaS) market is the fastest growing segment, with current projected revenue for 2019 of $38.9 Billion USD and forecast to grow at an average 25 percent annually rate, with projected revenues of $76.6 Billion USD by 2022[1]. Amazon Web Services (AWS), a provider of IaaS, comprises 41.5 percent of market share. Microsoft Azure and Google Cloud Platform (GCP) follow with 29.4 percent and 3.0 percent respectively [2].

2 Scope and size of the Spot price market and future trends

With the shift from IT being a capital expenditure to a managed service, companies must rethink how they approach utilizing and optimizing these resources, to maximize productivity and minimize costs. Cloud services have been widely touted as being less expensive than traditional on-site resources, which may not be true in all cases. Since cloud services are a pay-as-you-go delivery model, all services, both large (Virtual Machines) and small (Static IP addresses), are charged. If not properly managed, cloud service fees can become excessive and erase any cost benefits associated with cloud resourcing. With proper resource management, cloud resources can be instrumental in reducing computing expenses. It is important to note that most firms don't have a line item that discloses the amount that they spend on cloud computing. It is also presumed that firms with more than $25 Million in annual revenue would have separate pricing structures than those that we report in our research.

3 Why Are We Investigating AWS EC2 spot instances

AWS has become an integral part of many corporations' digital infrastructure. To meet the needs of such a diverse client base, AWS must offer sophisticated, robust services that are on-demand, latent, and dependable. AWS services have moved beyond the original business model of data storage and now offer a full range of cloud computing services. AWS is by far the leading provider of cloud services and offers some of the most mature solutions in the industry. AWS offers a pay-as-you-go approach for pricing for over 120 cloud services.

3.1 AWS EC2 Instance

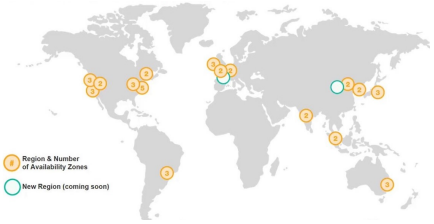
Amazon Elastic Compute Cloud (EC2) services are part of AWSs cloud computing platform which allows users to rent virtual cloud computing services (Amazon Machine Image (AIMs)), often referred to as instances, on which users can

[1] Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.5 Percent in 2019, 2019-04-02. Accessed: 2019-06-01. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g

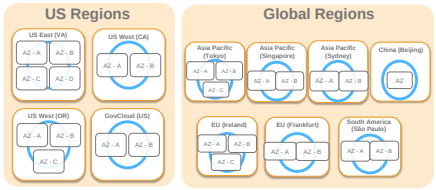
run/develop/deploy their own software/applications. User can create, launch, stop, hibernate, and terminate instances as needed, using a pay-as-you-go format, only paying for service while in use. EC2 instances offer three different resource pricing structures; On-Demand, Reserved and Spot instance pricing.

AWS's cloud services are distributed worldwide, to improve performance and reduce reliability. AWS's services are divided into Regions and Availability Zones.

- Regions: To improve fault-tolerance, decrease latency, and manage network traffic, AWS EC2 resources are strategically hosted at multiple locations worldwide, called Regions, Figure 1a. Currently, there are 20 regions (Americas-6, Asia-9, Europe-5).
- Availability Zones: Each region is further subdivided into isolated locations known as Availability Zones, Figure 1b. Availability Zones are independent of one another and do not share common infrastructure, adding an additional layer of fault tolerance. AWS allows users to specify the location of their EC2 instances (Region-Availability Zone) at time of launching. Example: ap-northeast-1c is an EC2 instance located in the northeast region in availability zone 1c.



(a) AWS Global Infrastructure



(b) AWS Region Makeup

Fig. 1: Amazon Web Services Regional Map and Availability Breakdown

4 AWS EC2 Pricing Structure: On-Demand, Reserved, Spot

AWS EC2 pricing structure is broken up into three (3) different models: On-Demand, Reserved, Spot pricing. Each model is tailored to take advantage of a users requirements and offer pricing solutions that best fit their needs.

- On-Demand Instance Pricing: Users pay for compute resources on a per-second basis, and pricing depends on the instance type, region, and availability zone. Just as the name implies, On-Demand instances do not require long-term commitments or up-front payments but come with the disadvantage of being higher priced. Instances are good for short-term applications,

unpredictable workloads that cannot be interrupted, and/or short-lived development/testing environments.

- Reserved Instance Pricing: Allow users to significantly reduce EC2 instance costs by pre-purchasing resources and capacity from AWS, using long-term contracts. Reserved contract tends to have 1- or 3-year term which is paid up-front or monthly. In exchange; hourly rates for EC2 instances are significantly lower (up to 75 percent) than On-Demand pricing. Reserved contracts are Region and instance type dependent and cannot be substituted for other instance types of Regions. Reserved instances are good for steady state applications, where long-term persistence is needed; i.e. Websites, applications where reserve capacity is needed. Reserved instances are just that, while you do pay an up-front fee, users are only billed when the instance is being utilized.
- Spot Instance Pricing: Allows the utilization of excess AWS EC2 computing capacity for up to 90 percent less than On-Demand pricing. Hourly pricing for Spot instances is called the Spot pricing. Spot price fluctuates with AWS computing resource utilization levels. AWS adjusts Spot price based on long-term supply and demand for computing instances (Similar to the price of a stocks/bonds on exchanges. To obtain a Spot instance, a user must submit a bid, price they are willing to pay for requested instance. If the bid price is above the current Spot price, and resources are available, then user is awarded an instance and runs until it is terminated by the user or the Spot price exceeds the bid price, at which time, the instance can be terminated/interrupted. AWS provides a Spot Instance Advisor service[2], to help users determine an optimal bid-price, or users can use proprietary algorithms and/or best judgment to establish maximum bid-pricing.

AWS recently implemented a hibernation status options for Spot instances in lieu of the original termination status. Previously, if Spot prices rose above a users spot instance maximum bid price, the instance would be reclaimed by AWS. Unless work was previously check-pointed[3], all work (data, computations) would be lost. This was a substantial inconvenience, risk, and reason why a user might not utilize spot price instances. Hibernation status essentially put the instance in a state of suspended animation. If a Spot instance is being reclaimed by AWS, hibernation allows users to save instance configuration, data, and any completed operations in RAM, which can be reloaded later. In the case of spot price instances, if AWS Spot prices rise above the users max bid price, an instance can be hibernated until such time that the current spot price fall below the maximum bid price, at which time, the instance can be restarted and previous operations can resume. Spot Instances are cost-effective if users are flexible about when your workload can be executed and can handle interruptions. Spot pricing

[2] Spot Instance Advisor, Accessed: 2019-06-03. [Online]. Available: https://aws.amazon.com/ec2/spot/instance-advisor/

is well suited for machine learning (ML) tasks, data analysis, batch processes, background activities, etc.

5 EC2 Spot Price Utilization Strategy

Cloud providers maintain large amounts of excess capacity to accommodate peak demands for their services; consequently, resources are frequently under-utilized. Since cloud resources are perishable commodities. To improve return-on-investment (ROI), providers have developed strategies to maximize utilization while minimizing idle capacity. One method is offering unused / excess computing resources at substantial discounts compared to other pricing tiers.

EC2 Spot pricing[4], determined by AWS, takes advantage of excess capacity on AWS's network and can be purchased at a substantially discounted prices. The business challenges are that spot prices fluctuate with network demand and instance can be terminated with minimal warning by AWS if network loading increases and/or current spot prices rise above the users maximum bid price. Taking advantage of Spot pricing, while maintaining an acceptable level of performance, are prudent objectives for any business.

Figure 2, depicts a hypothetical AWS EC2 Pricing Scenario showing typical On-Demand, Reserved, and Spot pricing data for an instance. We will step through several scenarios / strategies on how business can take advantage of the different pricing structures, reducing operating costs and maintaining performance levels.

As with most businesses, reducing costs are paramount. As we can see the hypothetical scenario in Figure 2, EC2 instance spot pricing are substantially less expensive than On-Demand pricing but they are not consistent, which introduces a new variable to the equation (variability/risk) when businesses try to exploit this opportunity. We investigate methods to reduce risk and increase predictability to businesses utilizing AWS EC2 Spot price instances by accurately predicting Spot instance pricing over a specified time-frame by using the advanced deep learning technique long short-term memory (LSTM) neural networks; thus, allowing businesses to determine optimal Max-Bid pricing, reducing instance termination risks and decreasing expenses.

In the following scenarios, Company A requires compute resources for an allotted time (time-frame A-F, in hrs.). These scenarios use the same instance type, Operating System, Time-frame, and AWS Availability Zone.

- **Scenario #1:** Company A utilizes On-Demand pricing for the required resource. In this scenario, they pay $0.78 * time-frame for the EC2 instance. No further payments or commitments are required.

[3] Checkpointing consists of saving a 'snapshot' of the application's state, so that it can restart from that point in case of failure.

[4] Announcing Amazon EC2 Spot Instances, 2009, Accessed: 2019-06-01. [Online]. Available: https://aws.amazon.com/about-aws/whats-new/2009/12/14/announcing-amazon-ec2-spot-instances/

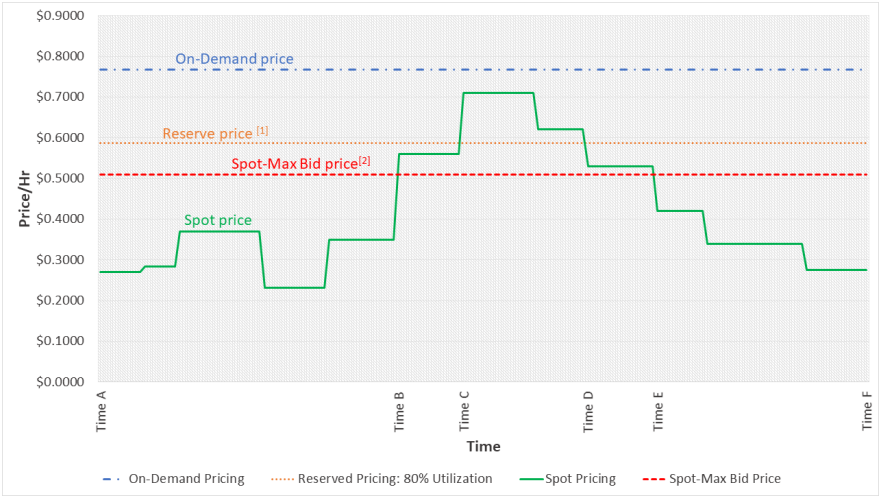


Fig. 2: Hypothetical AWS EC2 Instance Pricing Scenario. [1] Reserved price is an upfront yearly fee pricing plan for an EC2 Instance configuration. For this scenario, Reserved contract price is converted to an hourly fee, assuming 80 % utilization. [2] Spot-Max Bid Price - Maximum price user is willing to pay for this instance.

- **Scenario #2:** Company A utilizes a Reserved instance. Company A now pays an average of $0.60 * time-frame for the EC2 instance. Note that Reserve instances are pre-purchased for extended terms, so unless Company A anticipates requiring this same instance type repeated of the life of the contract (1 or 3 year terms), cost per compute hour will increase substantially. In the scenario in Figure 2, we assumed 80 percent utilization level for this instance.
- **Scenario #3:** Company A utilizes spot-Pricing for the required resource. To control budgeted expenses, Company A establishes a Max-Bid Price. Company A pays the much reduced, yet variable, Spot-price hourly rate from Time-A to Time-B. At Time-B, the Spot-price becomes greater than Max-Bid Price. At this time, either the instance is terminated and reclaimed by AWS or the instance is placed in Hibernation, depending on how Company A set up the instance. We will assume that Company A has chosen to hibernate the instance. At Time-E, the Spot price again moves below the Max Bid Price and the instance is restarted from the last checkpoint. From Time-E through Time-F, Company As instance is operational and being charged at the reduced Spot-price rate.

In Scenario #3, Company A dramatically reduced their overall cost for utilization of AWS EC2 instance over the other 2 scenarios. Being able to accurately predict EC2 instance Spot-pricing and setting proper Max-Bid pricing will en-

able Company A to improve financial results without negatively affecting their performance.

6 Related Work

AWSs EC2 spot pricing tier was launched back in 2009, with the discounted resource pricing structure of as much as 90 percent over on-demand pricing. Since that time there have been numerous research articles published related to different methodologies and techniques to capitalize on this EC2 Spot pricing tier. Techniques studied range from prediction, [spot price modeling/prediction], to managerial [task scheduling, resource allocations], to procedural [check pointing, AWS Spot fleets].

- **Technical Predictive Works:** Prior works focused on spot price modeling and predictions have implemented a range of methodologies. Since EC2 Spot pricing history displays many of the traditional characteristics of time series data, like what would be typical for a stock prices, researchers employed traditional auto-regressive / time series (ARIMA) modeling[3],[4]. Other researchers utilized more advanced techniques; e.g. Neural-Networks (NN) and long short-term memory (LSTM), deep-learning models [5],[6],[7]. Both types of approaches produced usable models with **reasonable prediction accuracy over a reasonable time window.**
  Other researchers focused their analysis on investigating spot pricing datas features; e.g volatility, feature importance, temporal trending[8],[9],[5]. The objective of these studies was to develop indices of the behavior of the pricing data, i.e. how volatile the pricing was, for users to be better informed when placing spot pricing orders regarding the past/future tendencies of the resource pricing. These techniques provide a better 'feeling' of the likelihood an instance may be terminated due to excessive price fluctuations.
  Nearly all the existing research predates a major change in the way Amazon manages the pricing model of the Spot price instances. In 2018 AWS launched a new pricing model, simplifying Spot pricing[5], reduced volatility, updated less frequently, and increased its predictability. Previously, AWS Spot pricing was driven by short-term trends in supply and demand; the new pricing model is based on long-term resource supply and demand trends, not higher competing bids. Due to this substantial shift in market conditions, researchers would be prudent to re-evaluate their pricing prediction models to see if they are adaptable to the new pricing model.
- **Resource Optimization Works:** Researchers studied other areas of resource allocations for improving utilization of AWS EC2 Spot market. Studies were conducted on novel approaches to resource provisioning by considering instance pricing, quantity of instances required, and task scheduling in order to minimize costs[10],[11].

[5] R. Pary,New Amazon EC2 Spot pricing model: Simplified purchasing without bidding and fewer interruptions, Mar 13 2018, Accessed: 2019-06-01. [Online]. Available: https://aws.amazon.com/blogs/compute/new-amazon-ec2-spot-pricing/

Several research papers looked at reducing costs by taking advantage of pricing inequalities between AWS regions[12],[13]. They conducted detailed analysis on pricing patterns of numerous AWS regions, to determine location effects on Spot instance pricing. Their research indeed showed that it was possible to lower overall costs by scheduling tasks in regions where Spot instance pricing was more attractive.
Spot Fleet[6]is a relatively new service offered by AWS. It is a collection of Spot Instances that are deployed to meet a clients compute instance capacity needs. If the Spot Instance Fleets allocation strategy is lowest Price, the Spot Fleet selects Spot Instances to fulfill the necessary capacity, using lowest cost instances available at the time. If any of the instances become unavailable, the Spot Fleet will automatically look for the lowest priced Spot Instance currently available, which may be of a different size or type, that would maintain capacity and add that instance to the Fleet. Spot Instances are by nature unstable, meaning they can be terminated with minimal notice.

– **Risk Mitigation Techniques:** Check pointing schemes, as to not lose all the work, are a legitimate risk reduction practice. Researchers studied optimal Checkpoint schemes that considered current vs bid Spot Pricing, time duration, application specific requirements, and other factors[14],[15]. Since Checkpoint add additional resource time and storage requirements, it is imperative that checkpoint be done only when needed to maintain an acceptable level of risk.

AWS EC2 Spot Instance price dataset utilized in this study was obtained directly from AWS and currently not hosted by any 3rd party organization; therefore, continued access to this data is determined by AWS policies and procedures. In the event that the pricing data become unavailable through AWS, we have found no pseudo/substitution pricing data publicly available from other sources that could be used in lieu of actual Spot Instance pricing data.

7 Case Studies - Firms utilizing AWS EC2 spot instances

– **Guttman Lab for lncRNA Biology**[7]
The benefits of using a Spot instances verses On-Demand pricing often can be the difference in being able to efficiently manage the expenses of start-up or keep cost in line with the research projects at universities throughout the world. In the case of the Guttman Lab at Caltech, Spot pricing allowed the research laboratory to reduce costs. Guttman studies genomic features called short for large non coding RNA (lncRNAs).
Expenses are a major concern for the lab, thus in addition to their research, it was imperative that they developed a cloud computing strategy that could

[6] How Spot Fleets Work, Accessed: 2019-06-01. [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/how-spot-instances-work.html

[7] Caltech Guttman Lab Case Study, Accessed: 2019-06-01. [Online]. Available: https://aws.amazon.com/solutions/case-studies/caltech-guttman-lab/

handle the capacity of their workload while managing expenses was developed. Working with Amazon Web Services, the Guttman Lab began to employ a strategy where they are using Spot prices as a part of their cloud computing ecosystem. The results are as much as a ninety percent costs savings as compared to on demand prices. These cost savings have a direct impact on allowing the lab to expand and increase work in gnomic research.

– **Gett**[8]
  Gett is an Israeli based start-up that is growing at more than three hundred percent per year. The company connects people with taxi drivers in more than fifty major cities throughout the world. In order to meet this massive growth, the firm must quickly scale its website and mobile back end to keep up with customer demand. As a start up, it is paramount to reduce costs to sustain growth and meet the demands of the firms investors.
  The firm runs its website and mobile app on several hundred AWS EC2 instances and approximately ninety percent of on the companys non-production EC2 instances use Spot pricing. Conversely, sixty five percent of the firms production instances use a Spot pricing strategy. The Spot pricing strategy employed by Gett has resulted in over $800,000 of annual savings.

Once an organization can accurately predict spot pricing, it could implement methodologies and techniques to capitalize utilization of AWS EC2 Spot pricing tier. There are numerous methodologies that can utilize AWS EC2 Spot Pricing predictions, here are just a few examples ways an organization might incorporate spot price prediction into their overall compute resource strategy.

– **Scheduling:** Organizations can incorporate spot pricing prediction into a resource/task scheduling matrix, to determine the optimal time to execute non-time-critical task. Utilizing the resources when pricing is most advantageous.
– **Location Determination:** EC2 Spot pricing varies by Region and Availability Zone. An organization can shop-around to locate the Region/Availability Zone where the required resource is predicted to have the lowest costs and obtain the required resources in that Region/Availability Zone.
– **Optimal Compute Configuration Determination:** Organizations can implement a strategy to optimize the compute resources required for a task based on predicted spot prices. For example; A task might require an EC2 instance type m5.xlarge [4 vCPU, 16GB] for 8 hours, after analyzing predicted spot instance pricing, it might be suggested that the organization utilize an m5.2xlarge [8 vCPU, 32GB] for 3 hours, based on overall total spot price predicted for the resource.
– **Strategic Checkpointing:** Strategic checkpointing is a risk reduction practice that can be implemented that considers current vs bid Spot Pricing, time duration, and other factors. Checkpointing adds additional resource time and storage requirements; therefore, checkpointing should only be used when risk

[8] "Gett Case Study," Accessed: 2019-06-01. [Online]. Available: https://aws.amazon.com/solutions/case-studies/gett/

levels are high enough to warrant the added expenses. For example; an organization might implement a rule that requires all tasks be checkpointed when the current EC2 sport price is within 10 percent of the bid Spot Price used to secure the resource.

8 Data Description

Spot pricing for Amazon computing instances, often called EC2 instances, is set by Amazon and adjusted based on supply and demand for compute instance capacity. AWS's Spot price database contains five (5) parameters: AvailabilityZone, InstanceType, Price, ProductDescription, and Timestamp. Table 1 shows an example of AWS's Spot price database.

Table 1: AWS Spot Pricing

Timestamp	Instance Type	Description	Availability Zone	Price
2019-05-08 21:46:36	c3.8xlarge	Windows	ap-northeast-1a	1.6503
2019-05-08 21:46:36	c3.8xlarge	Windows	ap-northeast-1c	1.7461
2019-05-08 21:46:34	i3.large	SUSE Linux	ap-northeast-1c	0.1223
2019-05-08 21:46:34	c4.8xlarge	Linux/UNIX	ap-northeast-1c	0.0223

AvailabilityZone: Amazon EC2 instances are hosted in multiple locations worldwide. Locations are comprised of Regions and Availability Zones. Regions are distributed geographically, placing them closer to users. Each region has multiple Availability Zones, distributed within that region. Amazon allows users to specify the location of their resources, in our case EC2 instances. An Availability Zone is represented by a region code followed by a letter identifier, Example us-east-1a. Compute resource (EC2) utilization levels, network bandwidth, time-of-day, can vary by Availability Zones, hence affecting EC2 Spot pricing.

InstanceType: Amazon EC2 provides numerous instance types optimized to for different user's use cases. Instance types comprise of different, predefined, combinations of CPU type/size/quantity, RAM, storage, and network capacity, see Table 2. This gives the user ability to choose the appropriate combination of resources for their needs. Be it high-bandwidth website hosting, compute intensive machine learning, or just general purpose, Amazon has a predefined instance type. Amazon has over 200 predefined Instance types available to users. Some are rarely used while others are used extensively.

Instance Type is represented by a 2 digit alphanumeric descriptor of the process followed by an description of the compute power of the instance, Example m4.large. Instance pricing is heavily dependent on the amount of resources dedicated to an instance. For example, on-demand pricing for t3.small, running Linux in us-east region, is $0.0208/hr., while pricing for m4.xlarge, running Linux in

Table 2: EC2 Instance Types

Model	vCPU	Memory(GiB)	Bandwidth(Mbps)	Network Performance
m4.large	2	8	450	Moderate
m4.xlarge	4	16	750	High
t3.small	2	2	not specified	Up to 5Gbps

us-east region, is \$0.20/hr. (9.6times)[9]. For instance types that are in high demand, Spot pricing tends to fluctuate more than would be expected for instance types with lesser demand. We will focus our efforts on the more readily used instance types for this study.

Price: Price is Amazon's EC2 instance Spot price, in US dollars, for an instance-type, given the availability zone and operating system. (ProductDescription), at the time specified (Timestamp). Pricing is not updated on a regular time interval but rather updated only when the Price changes, i.e. change in condition. The price for an instance is valid within the time period from the current entry (current timestamp) until a new entry is populated (next timestamp).

ProductDescription: Amazon EC2 has several options when it comes to operating system. Common options are Linux, SUSE Linux, Microsoft Windows, Windows Server, Ubuntu, and Red Hat Linux. Also available are pre-installed software licenses for commonly used utility software; SQL Server, PHP, Docker, Python, Java, *etc.*. Some software, such as Microsoft Windows, requires usage licensing. Pricing of instances is reflective of the additional costs associated with differences in the pre-installed OS and utility software described in the ProductDescription.

Timestamp: The time stamp of the Spot price (Price), is in UTC format (YYYY-MM-DDT*HH*:MM:SS Z) which is sometimes referred to 'Zulu time'. Example: 2019-03-31 07:28:29+00:00. The timestamp represents the time at which the spot price 'Price' of that EC2 instance-type, given the availability zone and operating system (ProductDescription), became valid.

Unlike financial stocks and bonds, whose prices are determined in financial markets, AWS EC2 instance Spot prices are based on supply and demand. However, AWS instance pricing is not a true open market where supply and demand are the key driving force behind pricing, but instead is a pseudo market driven pricing structure where AWS, using internal algorithms, determines spot pricing. These algorithms are based on instance supply and demand, network capacity, utilization levels, and other parameters determined by AWS to be a good representation of the current value of the asset (compute instance). The publicly available AWS EC2 instance Spot Pricing data set only contains the Spot price 'Price' determined by AWS algorithms for the particular asset.

[9] Amazon EC2 Pricing, Accessed: 2019-06-03. [Online]. Available: https://aws.amazon.com/ec2/pricing/on-demand/

AWS pricing algorithm and parameters used, computer and network utilization levels, supply and demand indicators, profit targets, peak demand pricing structure, day and time pricing structure (weekdays between 08:00 hours and 18:00 hours are premium time slots), special events (DDos, terrorist attacks), governmental requirements (security, taxation, reporting), and other factors (seasonal/holiday trends,) are proprietary and not publicly available. Since AWS' pricing algorithm is proprietary, we consider the Spot price as an agglomeration of all the relevant parameters when predicting AWS EC2 instance Spot prices.

9 Neural Networks and Deep Learning

Within the context of the Data Science world, neural networks could be defined as a set of algorithms that are loosely modeled after the human brain and, by design, they recognize certain patterns. Sensory data is interpreted through a machine perceptron that labels and clusters raw input. During this process there is translation of the data, whether it is text, images, or time series, into a numerical format which are contained in vectors.

Neural networks can be used for reinforcement learning, classification and regression and are often described as a clustering and classification layer on top of data that you store and manage. Figure 3 shows an example of a Neural Network. The network will receive a certain number inputs that are fed to a weight that represents the strength of the connection. The weight has the ability to bring down the importance of the input value. Activation functions modify the data prior to sending the impulse to the output layer. The activation function is what allows neural networks to model complex, non-linear relationships.

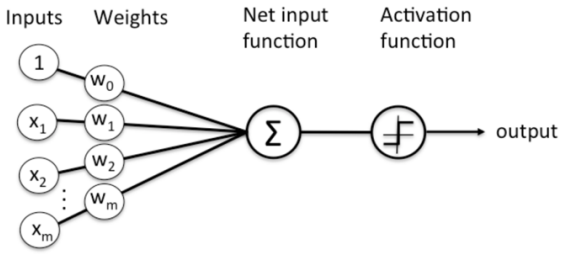


Fig. 3: Neural Networks

Deep learning is a sub field of machine learning that was inspired by the neural networks. It is best described as a large, deep neural net. Specifically, deep learning refers to the number of hidden layers in a neural networks that are used as part of deep learning algorithm that could be used in image recognition, natural language processing, and recommender system problems.

10 Autoregressive Integrated Moving Average (ARIMA)

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used for analyzing and forecasting time series data. It is a generalization of the AutoRegressive Moving Average with the added feature of integration.

The ARIMA model can be broken down as follows:

- **AR:** Autoregression
- **I:** Integrated
- **MA:** Moving Average

The parameters of the ARIMA model are defined as follows:

- **p:** The number of lag observations included in the model, also called the lag order.
- **d:** The number of times that the raw observations are differenced, also called the degree of differencing.
- **q:** The size of the moving average window, also called the order of moving average.

The selection of appropriate values for p, d, and q depends on the particular time series. Generally speaking, longer training windows and shorter forecast windows lead to a more accurate forecast.

11 Long Short Term Memory networks (LSTM)

Long Short Term Memory networks (LSTM) LSTMs are a particular kind of Recurrent Neural Network (RNN) that are capable of learning long-term dependencies. What sets an LSTM network apart from a RNN is its ability to avoid the long-term dependency problem. By design, an LSTM network behavior remembers information for long periods of time. RNN have difficulty with this type of behavior. LSTM is the solution for remembering the gradient values during backpropagation, as it replaces the RNN cell.

The equations below demonstrate how updates are less chaotic in accounting for all of the learning that brings order to all of the memories that the gate is responsible for learning. LSTM is a more extensive series of matrix optimization. Each gate in the LSTM is a series of matrix operations and performs like a main neural net where you can compute the derivative of each component.

$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right) \tag{1}$$

$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right) \tag{2}$$

$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right) \tag{4}$$

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right) \tag{5}$$

$$h_t = \tanh(C_t) * o_t \tag{6}$$

The main idea behind LSTMs is the cell state, which is a horizontal line that runs through the top of the diagram below. The LSTM has the ability to add or remove information to the cell state. The cell states are regulated by structure that we refer to as gates.

LSTM networks hold a particular advantage over RNNs due to the three internal gates that allow the LSTM to forget data. In other words, excess data that is filtered through a RNN during backpropagation can clutter up relevant information and allow it to be lost in the process. The nature of the setup of the LSTM and the gate that allows the network to forget provides a strategic advantage over conventional RNN.

In a time-series model, LSTM networks can be extremely effective in predicting an arbitrary number of steps into the future. There are five essential components that allow the LSTM network the ability to model long-term and short-term data. However, RNNs repeatedly apply the same network to themselves. This turns out to be the tragic flaw of RNNs and one of the major factors where LSTM networks have a major advantage.

- **Cell state:** The internal memory of the cell that stores short and long term memory.
- **Hidden state:** An output state information that calculates current input, previous hidden state and current cell input that will eventually predict the future movement of the spot price.
- **Input gate:** Determines how much information from current input flows to the cell state.
- **Forget gate:** Determines how much information from the current input and the previous cell output will flow into the current cell state.
- **Output gate:** Used to decide how much information from the current cell state flows into the hidden state. This allows the LSTM to pick long or show term memories.

One of criticisms of the LSTM architecture is that many of the components that make up the network can be difficult to understand and in some cases ad-hoc. With any good analysis we should keep an open mind and understand that it is possible that better architectures exist.

12 LSTM versus ARIMA

By its very nature, one would correctly assume that deep learning work flows will demand more computing capacity for models to test and train. AWS Spot pricing shows that the GPU instances are among the most expensive of the available spot prices and the ones that lend itself best to deep learning. We will also seek to prove that deep learning models that use GPU are most advantageous when you surmise your computing needs and compare it to the spot instance discount that you receive.

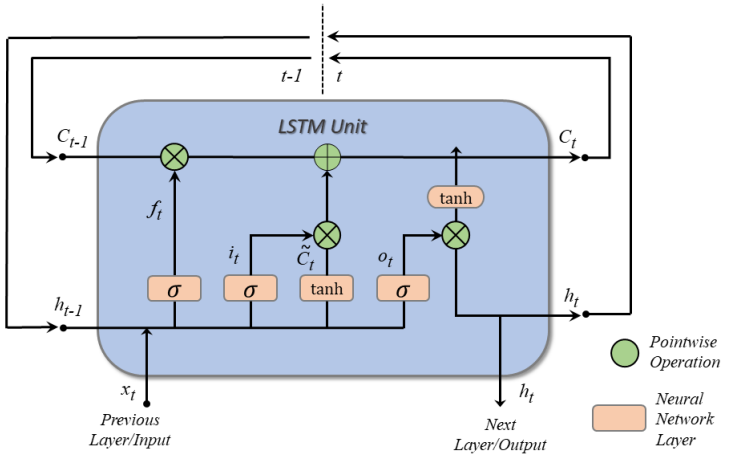


Fig. 4: LSTM Model

AWS provides a list of solutions to deal with the two-minute time notice. The intent of this advice will allow the user to be prepared to deal with the potential loss of data, while still being able to take advantage of a Spot pricing strategy. Some of the suggestions are to decouple compute, storage and code artifacts, and keep the compute instances stateless, use a dedicated volume for datasets and training progress, and to automate replacement instance creation after termination [14].

13 Evaluation

We were able to accomplish our goal of determining spot price direction by using a sample set of the data. This was computationally less expensive, while maintaining the integrity of traditional statistical sampling.

LSTM networks are designed to address the problem of long-term dependency. Defining the model to predict the difference in values between time steps rather than the value itself, is a much stronger test of the models predictive powers. Our model uses a simple three-layer network composed on two LSTM layers. Each layer is 32 units wide. Finally, we have a dense node to consolidate input form the second LSTM layers to determine the final value.

To train our neural network, we use the ADAM optimization algorithm and Mean Absolute Percentage Error (MAPE) as our loss function. We ran a total of 20 epochs to train our model. The results showed the LSTM Networks with an average reduction in MAPE of approximately 95 percent, compared to the baseline ARIMA model. Table 4 shows an example of the evaluation metrics that we utilized to measure the effectiveness or our model.

The central idea of this analysis is to optimize spot price predictions and allow a business to enjoy the best possible price to pay for cloud computing.

Table 3: Dataset Description

Attribute	**Description**
Source	AWS API
Description	AWS Spot Instance Pricing
Region	US East
Availability Zone	us-east-1b
Operating System	Linux
Variables	AvailabilityZone, InstanceType, Price, ProductDescription, Timestamp
Start Date	03-31-2019
End Date	05-30-2019
Number of Observations	1,414
Size of Dataset	40 MB

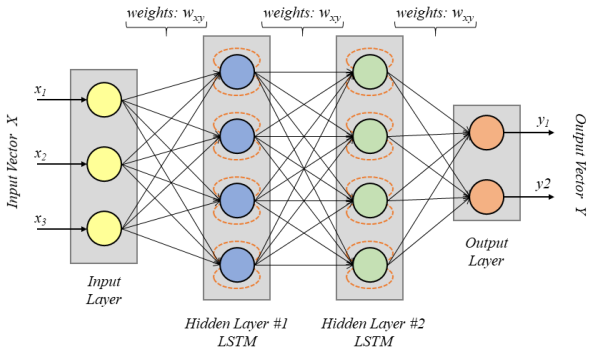


Fig. 5: Neural Network LSTM

This is obviously appealing to any business, but we have to remember the two-minute cutoff that AWS has the right to enact. Thus, if you are a high frequency hedge fund, you will not be able to take advantage of these steep discounts.

The LSTM network is designed to process tick data. Figure 6 below shows the results of our model and considers a total of 300 ticks (roughly 25 hours) as the time measurement. The chart include the Region, Availability Zone, Description and Spot Price and Time-series. Notice that the model prediction is well within the band of the actual Spot price.

14 Ethics

Several ethical and business issues should be addressed prior to implementing a strategy that incorporates the use of AWS EC2 Spot instance forecasting. Access to Data: AWS EC2 Spot Instance price dataset utilized in this study was obtained directly from AWS. Currently, AWS provides this data freely through

Table 4: Model Evaluation Metrics

Metric	Formula	Type	Observations
Mean Absolute Percent Error (MAPE)	$MAPE = \text{mean}\left(\left\lvert\frac{e_t}{y_t}\right\rvert\right)$	Percentage	Undefined if some $y_t = 0$
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE}$	Scale-dependent	Similar to standard deviation

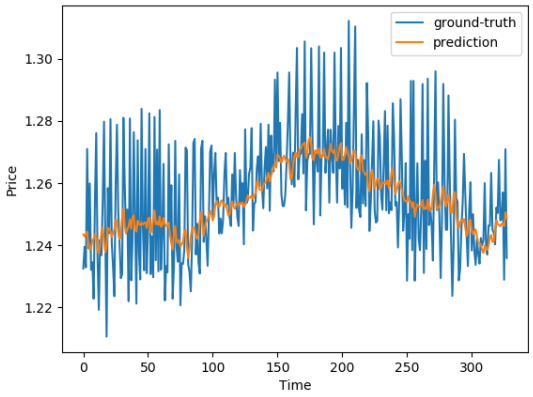


Fig. 6: LSTM Network Result

easily invoke APIs or directly from their website[10]. The price data is updated real-time. This makes obtaining new data trivial and the prediction models can be updated frequently.

The EC2 Spot pricing data is not hosted by any 3rd party organization; therefore, access to the Spot instance pricing data is determined solely by AWS policies and procedures. Given AWS hyper-consumerist behavior, policies regarding access to this pricing data could be changed at any time. As more customers become increasing aware of Spot instance pricing and how it may benefit their business by reducing cost, AWS may seek to change their policies in a manner that will increase their own profitability.

Data Access:

Several scenarios could be envisioned.

– Data access remains free but is delayed. This would make predicting pricing more difficult and possibly less accurate. This is similar to how financial stock-data is currently distributed.

[10] Amazon EC2 Pricing, Accessed: 2019-06-03. [Online]. Available: https://aws.amazon.com/ec2/pricing/on-demand/

– Access to data becomes controlled and/or monetized. If numerous customer begin utilizing the pricing data, AWS may see this as a new revenue stream, which can be monetized to increase profits and/or recover some lost-profitability that will come from the increased utilization of Spot instances vs Reserved/On-Demand instances.
– AWS impart policies and procedures to limit the use of Spot price instance data. The data is currently available to consumers, to assist them in making decisions about usage of Spot instances. If the usage of this data changes, to larger scale industrial type of applications, restrictions on utilization could be implemented.

Since Spot instance prices are developed internally within AWS, a secondary source of data is unavailable if AWS ceases to distribute the data. All Spot instance pricing features are based upon are not public knowledge, therefore, there is no viable alternative source for pricing data and/or data that might be able to substitute for Spot pricing (demand, network traffic, etc.).

Data Integrity: The Spot instance pricing is based on proprietary algorithms based on several utilization factors (instance demand, network traffic, resource availability, etc.) and cannot be independently verified. The pricing is, in essence, a black-box algorithm and not based on open-market style forces. While we hope that all companies are fiduciary with the data they present to the public, there is no laws that require data to be accurate.

While AWS is currently openly sharing Spot instance pricing, this does not mean policies/procedures might change in the future. Because of this, consumer must be cognizant of the limitations and potential negative effects that might occur in the future and should plan accordingly when implementing a strategy that utilizes Spot instance pricing predictions.

15 Conclusion

Cloud computing has changed the way that companies access computing resources to run their business. This study investigates methods to reduce risk and increase predictability of pricing for businesses utilizing Amazon Web Services(AWS) elastic compute cloud(EC2) Spot instance pricing tier by accurately predicting Spot instance pricing over a specified time-frame using Long short-term memory (LSTM) neural networks and comparing the results against traditional time-series auto-regressive integrated moving average (ARIMA) modeling. Spot instances pricing can be up to ninety percent less than typical On-Demand pricing, providing organizations significant savings. If organizations are able to accurately predict Spot instance pricing, thus allowing them to implement sourcing strategies based on the predicted pricing, substantial compute resource costs savings can be realized.

AWS EC2 instance pricing is broken down into three tiers; On-Demand, Reserved and Spot. On-Demand instances do not require long-term commitments or up-front payments but come with the disadvantage of being higher priced. On-Demand instances are appropriate for short-term, unpredictable workloads

that cannot be interrupted. Reserved Instance pricing significantly reduces EC2 instance costs, using long-term contracts. Reserved contract tends to have multi-year terms. In exchange; Reserved Instance hourly rates for EC2 instances are significantly lower (up to 75 percent) than On-Demand pricing. Spot pricing, the focus of our research, allows users to utilize excess AWS EC2 computing capacity for up to 90 percent less than On-Demand pricing. The business challenges are that Spot Instance prices fluctuate with network demand and instance can be terminated with minimal warning by AWS if network loading increases and/or current spot prices rise above the users maximum bid price. Taking advantage of spot pricing, while maintaining an acceptable level of performance, are prudent objectives for any business.

Traditionally, predictions of serially correlated data set, like the Spot Instance pricing data, has been accomplished using ARIMA modeling techniques. This sort of analysis can also be analyzed using Machine Learning or Deep Learning. When deciding which type of analysis to conduct, it is important to consider the size, shape and content of the data set being analyzed. The size and shape of the Spot Instance data set, few variables with numerous observations; i.e. skinny & long matrix, tend to be good candidates for using Deep Learning techniques. In this study, we utilized LSTM neural networks for our analysis. LSTM was chosen for its ability to; a) extract patterns over long observation time frames, b) manipulate its memory state, by selectively remembering patterns while forgetting non-essential information, and c) model both quantitative and categorical variables.

For comparison, we used an ARIMA model as baseline comparison for our LSTM model prediction accuracy. The result showed of the Spot Instance price predictions from our LSTM model has as an average reduction in mean absolute percent error (MAPE) of approximately 95 percent when compared to the baseline ARIMA model. While Deep Learning models can produce improved results, the models can be less robust and results can be difficult to interpret.

References

1. Mell, P., Grance, T., et al.: The nist definition of cloud computing. (2011)
2. : Custom applications and iaas trends 2017 (2018-04. McAfee, Santa Clara CA, 95054) Accessed: 2019-06-01.
3. Zhao, H., Pan, M., Liu, X., Li, X., Fang, Y.: Optimal resource rental planning for elastic applications in cloud market. In: 2012 IEEE 26th International Parallel and Distributed Processing Symposium, IEEE (2012) 808–819
4. Chhetri, M.B., Lumpe, M., Vo, Q.B., Kowalczyk, R.: On forecasting amazon ec2 spot prices using time-series decomposition with hybrid look-backs. In: 2017 IEEE International Conference on Edge Computing (EDGE), IEEE (2017) 158–165
5. Wallace, R.M., Turchenko, V., Sheikhalishahi, M., Turchenko, I., Shults, V., Vazquez-Poletti, J.L., Grandinetti, L.: Applications of neural-based spot market prediction for cloud computing. In: 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS). Volume 2., IEEE (2013) 710–716

6. Sarah, A., Lee, K., Kim, H.: Lstm model to forecast time series for ec2 cloud price. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), IEEE (2018) 1085–1088
7. Al-Theiabat, H., Al-Ayyoub, M., Alsmirat, M., Aldwair, M.: A deep learning approach for amazon ec2 spot price prediction. In: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), IEEE (2018) 1–5
8. Wang, C., Liang, Q., Urgaonkar, B.: An empirical analysis of amazon ec2 spot instance features affecting cost-effective resource procurement. ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS) **3**(2) (2018) 6
9. Lumpe, M., Chhetri, M.B., Vo, Q.B., Kowalczyk, R.: On estimating minimum bids for amazon ec2 spot instances. In: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, IEEE Press (2017) 391–400
10. Zhou, A.C., He, B., Liu, C.: Monetary cost optimizations for hosting workflow-as-a-service in iaas clouds. IEEE Transactions on Cloud Computing **4**(1) (2015) 34–48
11. Bhise, V.K., Mali, A.S.: Ec2 instance provisioning for cost optimization. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE (2013) 1891–1895
12. Lee, K., Son, M.: Deepspotcloud: leveraging cross-region gpu spot instances for deep learning. In: 2017 IEEE 10th International Conference on Cloud Computing (CLOUD), IEEE (2017) 98–105
13. Ekwe-Ekwe, N., Barker, A.: Location, location, location: exploring amazon ec2 spot instance pricing across geographical regions. In: 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), IEEE (2018) 370–373
14. Khatua, S., Mukherjee, N.: A novel checkpointing scheme for amazon ec2 spot instances. In: 2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, IEEE (2013) 180–181
15. Jangjaimon, I., Tzeng, N.F.: Effective cost reduction for elastic clouds under spot instance pricing through adaptive checkpointing. IEEE Transactions on Computers **64**(2) (2013) 396–409