AWS COST PREDICTION USING MACHINE LEARNING

Contents

1.0 INTRODUCTION

This article talks about predicting cost of AWS public cloud service usage in future, applying machine learning techniques with the input from AWS Cost explorer API.

It has two portions

1) Retrieve Past Cost information (Data) from AWS using AWS Cost Explorer
2) Use the cost data and apply Machine Learning Techniques for future Predictions

2.0 PRE-REQUISITES

1) Retrieve Past Cost information (Data) from AWS using AWS Cost Explorer
   a. AWS super user account to retrieve AWS Billing
   b. AWS secret key and Token for the super user
   c. REST API client (Node JS /HTML) to connect to AWS Cost Explorer API endpoint(https://ce.us-east-1.amazonaws.com)
   d. Database to store the Cost information retrieved in Json Format

2) Use the cost data and apply Machine Learning Techniques for future Predictions
   a. Database with AWS Billing history (One Year) -Mongo DB/MySQL
   b. Python Machine Learning Library (Sci Kit Learn)
   c. Numpy and Pandas Library to be used in Sci Kit

3.0 SUMMARY

Need for AWS Cost Prediction for the AWS Service usage

1) On day to day basis, Organisation/LOB/Project is using AWS public cloud service like AWS EC2 instance, S3, Elastic load balancer and other services which are required and configured for your project to run
2) Each service is billed per hour/month/sec based on billing options provided by AWS
3) At the end of each month AWS service usage billing can be downloaded and verified against usage on daily basis
4) As an organization, Project teams are facing challenges in controlling and prevent cost charges with or without notice due to
   a. Due to overlook of AWS Services and their charges
   b. Threshold set for each project has crossed due to unavoidable incident or changes
   c. Project Budget cost increase due to unnoticed AWS service cost usage

d. Cost predictions on top of automation and auto shut down of services due to unusual behaviour

e. Need to monitor and behaviour or various services and usage with the help of cost and usage

f. For Management review and analysis of cost usage data

How does Machine Learning help in predictions of AWS costs?

- Machine learning has options to predict the target value using previous data history using statistical formulae which can be predicted without

- Machine learning can be divided into
  - Supervised Learning
    - Classifieds
    - Regression
  - Unsupervised Learning

Note (for more information, please check the published documents related to Machine Learning)

- Supervised learning applies where you have history of data
  (In this case we have history of cost from AWS billing from cost explorer
    - Daily
    - Monthly
    - Last 7 days to Last One Year

Cost in the format of Json with Service Name

Sample

{"ResultsByTime": [{"TimePeriod":{"Start":"2018-01-01","End":"2018-01-02"},"Total":{"BlendedCost":{"Amount":"51.0605314896","Unit":"USD"}}

Whereas in unsupervised learning you have patterns and data but no pointers

Or information to map the data to any specific target value or features.

- In supervised learning we need to select
  - Regression – since we have continuous independent variable which is cost and other dependant variable are AWS Service name, AWS cost by Service, Day, month and year

newt|global
Aiming At Customer Delight

Technical challenges and limitations in this approach

- History of data (One-year data may not be enough in certain cases)
- Prediction precision is very important
- Margin of error is dependant on accuracy of data and machine learning algorithm
- Also, dependant on Machine learning train and test steps
- No of features or various other factors like Usage of service, Manual Errors Unusual incidents can mislead or make predictions not practical

How do we increase the accuracy of prediction by Narrowing Errors?

- Train the machine with as much as random past data to ensure the straight-line graph is achieved
- Try to predict the cost with various features or factors and not with single feature
    - For e.g. if your target value is Blended cost of AWS service per month then your
      Dependant variable should not only have blended cost of all service monthly or daily, it needs to have each service cost usage, Inventory list, dependency for this project, Incidents, Customer usage pattern, etc as appropriately

High Level Machine Learning Steps for future AWS Cost Predictions

1) Import data from AWS ->AWS Cost Explorer->Database->Sci Kit using Numpy and pandas
2) Map the history values along with features(column) to Python Sci kit
3) Assign y and x axis for linear regression algorithm to work and predict the mean and other statistical formulae using base equation y=MX =C

    Note: Sci kit learn provides functions and libraries for the same

4) Train and test the prediction of future date value using Python Sci Kit Learn libraries
5) Use R2 statistical algorithm and other techniques until the prediction value is with less margin of error

Note: Have skipped other programmatic and statistical steps for the simplicity of presentation

newt|global
Aiming At Customer Delight

5.0 CONCLUSION

In conclusion this approach will be help for AWS customers to proactively reduce the usage cost and facilitate management decisions. However, the accuracy and margin of error depends on training the machine, choosing the right model and the various factors provided as input and dependant variable for algorithm to arrive at the target value.