BACHELOR OF SCIENCE HONORS DEGREE IN SOFTWARE ENGINEERING

FINAL YEAR RESEARCH PROJECT PROGRESS REPORT

FACULTY OF SCIENCE

UNIVERSITY OF KELANIYA

Bi-Weekly research progress report submitted by the student

Student No: SE/2015/025
Student name: M.S.Prasad
Name of the research project: Cloud Service Selection Using Machine Learning
Name of the research supervisor: DR. LANKESHWARA MUNASINGHE
Report No.: 01
Period covered (from ~ to dates): 2020/10/26 – 2020/11/09

Instructions:

- Bi-weekly report should be completed in every two weeks and must submit to the research supervisor no later than 11:59 pm on the Friday of the second week in review.
- Provide descriptive answers for each of the progress review questions. You may spend 100 ~ 150 words for your answer (it must contain at least 50 words).
- You are strongly advised to contact your supervisor in every two weeks. In your report, you are required to state how you addressed the supervisor's comments for the previous report.

(1) State the overall research progress (from start to UpToDate).

After completing the literature review, I start to preprocess my datasets using ML algorithm. I have to do lot of preprocessing task to get valuable, cleansing dataset because that dataset is dirty. Firstly, I followed AWS General purpose dataset for analysis. It has many of null values, different type of data types, outliers, and like issues. So, I have to get lot of time to make dataset using pandas like python toolkit. As a example, when see the null values, if null values are low in that column that remove entire row (PricePerUnit column). And most of times, I add most frequency value instead of null values (LeaseContractLength – 3, PurchaseOption - All Upfront, OfferingClass – standard, Location- US West, Current_Generation-Yes) like that. In EffectiveDates date column, some values are string type, some Date format. So, I have to change

all string type to Date type. As well as some values are not usage in current time, that means some instant type removed by AWS cloud provider. So, I changed that types are current types. Some values are change to common type, because I want to extract type for some values (EBS only => EBS, *SSD => SSD, *HDD => HDD). Grouping datasets column values to comfort for using. There are Memory [(1-2), (2-4), (4-8), (8-16), (16-32), (32-64), (64-96), (96-128), (128-256), (256-384)], vCPU [(1-2), (2-4), (4-8), (8-10), (12-16), (16-24), (24-32), (32-48), (48-64), (64-72), (72-96)]. Adding new column as requirement in order to obtain user input. There are the Project Type - according to Instance type, vCPU, Memory and AWS recommended Use Cases and Project Size - according to Instance Type. Finally, I can get perfect dataset for my machine learning task. After I label all of columns, I could start my first algorithm testing. I used Project Type, Size, Memory, vCPUs, Storage and Network Performance as independent variable. I can use this for other Machine family type like these. I used Multiclass Classification with Support Vector Machines (SVM) to recognize to instance type for customer project. SVM has 5 kernel types for training the dataset. There are linear, poly, rbf, sigmoid, and precomputed. I get ready for comparing all of type algorithms to find and select best one approach.

(2) What were the supervisor's comments on the previous report and how did you address them?

I understood some writing issues in abstract and literature. So, I fix that issues like citation are not using in abstract. I used one citation in abstract. According to given comments, I can note down some main points to success my research project perfectly. There are flow of story, well organized progress week by week, nicely organized content like that.

(3) State the progress of your research compared to the previous two weeks period.

Previously, I found datasets, ml methods and other researcher's methods for selecting best cloud service provider. I followed 25+ research for doing this. They introduce many of methods for different purposes. I gathered all of ideas, methods, issues of their methods, limitations like that. After I planned to create start ml prediction with preprocessing dataset. I hoped to check it for General purpose family type in AWS EC2. As I aspect, I can do preprocessing task and start to predict values.

(4) What is your plan for next two weeks?

I plan to compare accuracy of ml learning algorithms to predict best result. And I hope to start my second prediction part of pricing calculation using linear regression algorithms. I have prepared more addition columns in dataset for this prediction. I will use Multiple linear regression. because, we have multiple variables like Location, Instance Family, Storage, Network. It can use all other service providers like this method.

(5) Any other matters related to your research.

I want to find dataset for Azure and GCP. I try to send emails to this cloud providers. But I cannot get opportunity for getting datasets to my research. I would like to know, if I cannot find other service provider's datasets, is AWS EC2 dataset is enough to show my evaluation?