**Capstone Project** Samar Shaaban Abdelfattah Haytamy Machine Learning Engineer Nanodegree February 15, 2019

Definition

Project Overview

The cloud computing idea is based on reusability of IT capabilities. The enterprises as well as individuals can use these Cloud-based services as an partial solution to their operational and business problems. The leading cloud computing providers(Google, Microsoft, Amazon, E-Bay, IBM, etc) have built An online marketplaces to facilitate the publication and searching of different types of cloud services in a more suitable way. The marketplaces provide services on demand, paying per usage and managing automatic service elasticity to meet users’ requirements (Cloud services).

The Cloud consumer usually needs to use Cloud services as a partial solution to his requirements. So, the appropriate Cloud services have been composed and provided as a single virtual service to the Cloud consumers.

In this project, I create a desktop application that capable of composing multiple services from different cloud providers (e.g. IBM, Google, Rackspace,). The application uses a predictor trained using QoS (quality of services) dataset[[1]](#footnote-1) to predict future provision values to accurately select providers.

The project was inspired by this paper[[2]](#footnote-2) (Z. Ye, 2016).

Problem Statement

The goal is to create a service composition application capable of recommending best providers to contract with; the tasks involved are the following:

1. Download and preprocess the QoS attributes data (provider side).
2. Generate the consumer preferences/services weights (consumer side).
3. Train predictor1 that can predict future provisioned QoS values(provider side)
4. Train predictor2 that can predict future QoS weights(consumer side)
5. Develop the composer (broker) application
6. Rank or recommend the best cloud providers.

The final application is expected to be useful for cloud end users (consumers) because it help him to contract with the best providers appropriate to his requirements.

Metrics

Root Mean Square Error (RMSE): RMSE is used to evaluate the performance of a prediction model.

Let assume that the time series of an individual attribute in the QoS history fits by many prediction models. If the predicted time series are ( ), the prediction error is calculated using the following equation. A lesser value of RMSE imposes a better prediction model.

RMSE (i) =

Analysis

Data Exploration

* A synthetic dataset is used to represent the cloud consumer preferences or his requirements which are his preferred weights for response time, throughput, availability for the required service :
* It is a historical time series data follow Gaussian distribution with ‘matern’ kernel. It is generated using the TimeSynth open source library (https://github.com/TimeSynth/TimeSynth ).
* The end user preferences data will be 3 series of floating point values of length 1440

Data Shape= 1440 rows × 3 columns

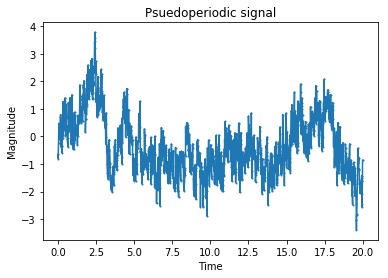
Note:

* The history of four years is 1440=360\*4.
* The interested attributes (3 columns/series); throughput, response time, availability.
* The cloud providers’ data set will be represented using a real cloud service data (W. Jiang, 2012) which is updated in (Haytamy S.S.). It contains 5 historical time series for 100 cloud service providers collected through 6 months as 28 time slots as follows: 1. Availability 2. Max Response time 3. Min Response time 4. Avg. Response time 5. Throughput.
  + - The values are floating point numbers
    - It is available here: (https://github.com/SamarShabanCS/Math\_for\_ML/tree/master/time%20series%20data%20QoS)

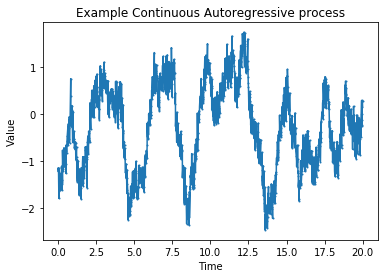
Exploratory Visualization

1. The cloud consumer preferences are synthetic data generated using TimeSynth library. The following graphs are availability, response time and throughput time series.

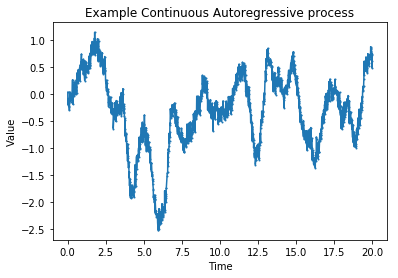
* Availability attribute



* Response time attribute

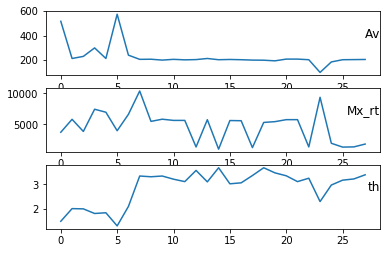


* throughput attribute



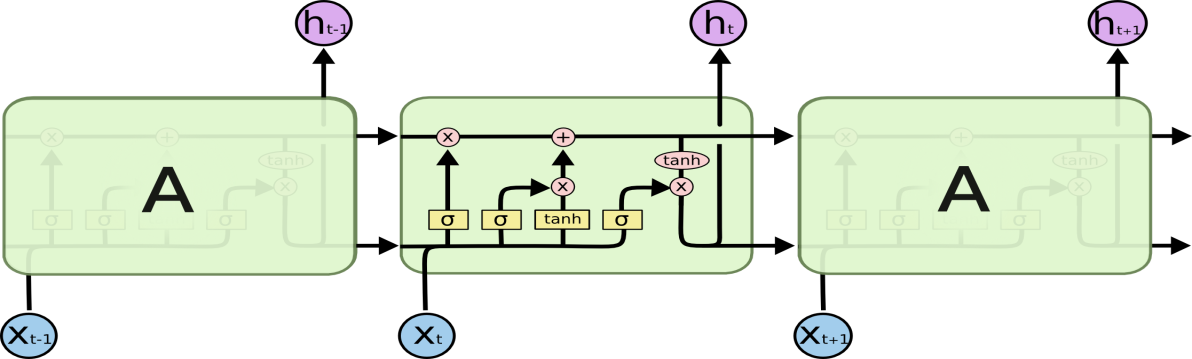
1. For provider QoS data, the following plot show the distribution or history of the QoS attributes (Av: availability, Mx\_rt: maximum response time, Th: throughput) provisioned from one provider selected randomly from the cloud providers’ data set. It shows that there is strong negative correlation ship between response time and throughput, when the response time is slower; the throughput is higher and vice versa. When calculating the correlation operator between these two parameters, it is found it is equal to -0.67 which ensure the mentioned description. This information helps us to use machine learning model that is capable of capturing this relation between Th and Rt.

Also, the QoS values are varied, so they need to preprocessed (scaling).



Algorithms and Techniques

The predictor is the LSTM [[3]](#footnote-3)(long short term memory) which is a type of recurrent neural network capable of learning order dependence in sequence prediction problems (Time series in our problem). LSTMs are designed to avoid the long-term dependency problem which is in traditional RNN. Remembering information for long periods of time is practically their default behavior. LSTms are illustrated in the following figure:



The following parameters can be tuned to optimize the predictor:

1. Input Parameters:

* Preprocessing and Normalization

1. Neural Network Architecture:

* Model Type (MLP or LSTM; mostly focused on LSTM)
* Number of Layers (how many layers of nodes in the model)
* Number of Nodes (how many nodes per layer)

1. Training Parameters:

* Training / Test Split (how much of dataset to train versus test model on
* Look back (how many prior days are included in the input sequence
* Batch Size (how many time steps to include during a single training step
* Optimizer Function (which function to optimize by minimizing error
* Epochs (how many times to run through the training process

Benchmark

According to my research, the nearest and closest paper to the mentioned problem is “Long-Term QoS-Aware Cloud Service Composition Using Multivariate Time Series Analysis” (Z. Ye, 2016) (Z. Ye, 2016). So, it will be used as the benchmark model. It propose a multiple QoS prediction model (MQPM) by using the Arima model to predict the QoS values then compose the services that match cloud consumer requirements by using the Euclidian distance.

The performance of MQPM of its predicted QoS attributes:

|  |  |  |
| --- | --- | --- |
| Prediction model | RMSE | |
| Avg. throughput | Avg. response time |
| MQPM | 0.32 | 59 |

Methodology

Data Preprocessing

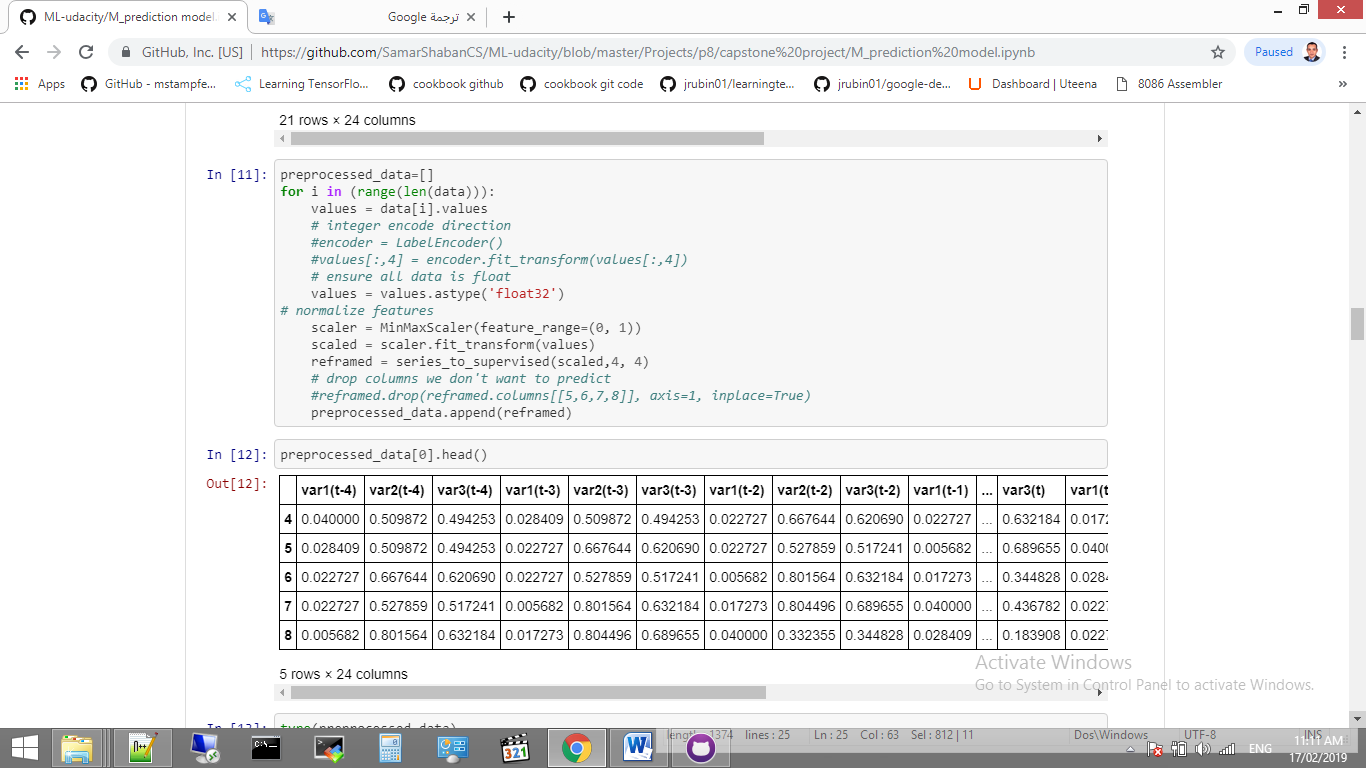
The following steps are done in the preprocessing phase before building and training the prediction model:

1. Generate the user QoS preferences values using TimeSynth generator. This was done in a Jupyter notebooks (titled “END\_user\_preferences.ipynb”
2. Read the QoS history.
3. put the data points into a Pandas DataFrame for ease of organization and visualization
4. Define train test split ratio
5. Create the training and test datasets
6. Define Look back (how many prior days to include at each time step)
7. convert from series form to supervised form(X,Y)
8. Normalize data (normalize 0.0 to 1.0 for better performance)

* Steps from 2 to 8 are done for both consumer data and provider data in the following notebooks:

(end\_user\_prediction model .ipynb, [M\_prediction model.ipynb](https://github.com/SamarShabanCS/ML-udacity/blob/master/Projects/p8/capstone%20project/M_prediction%20model.ipynb))

The following graph show sample of preprocessed data (from M\_prediction model.ipynb)



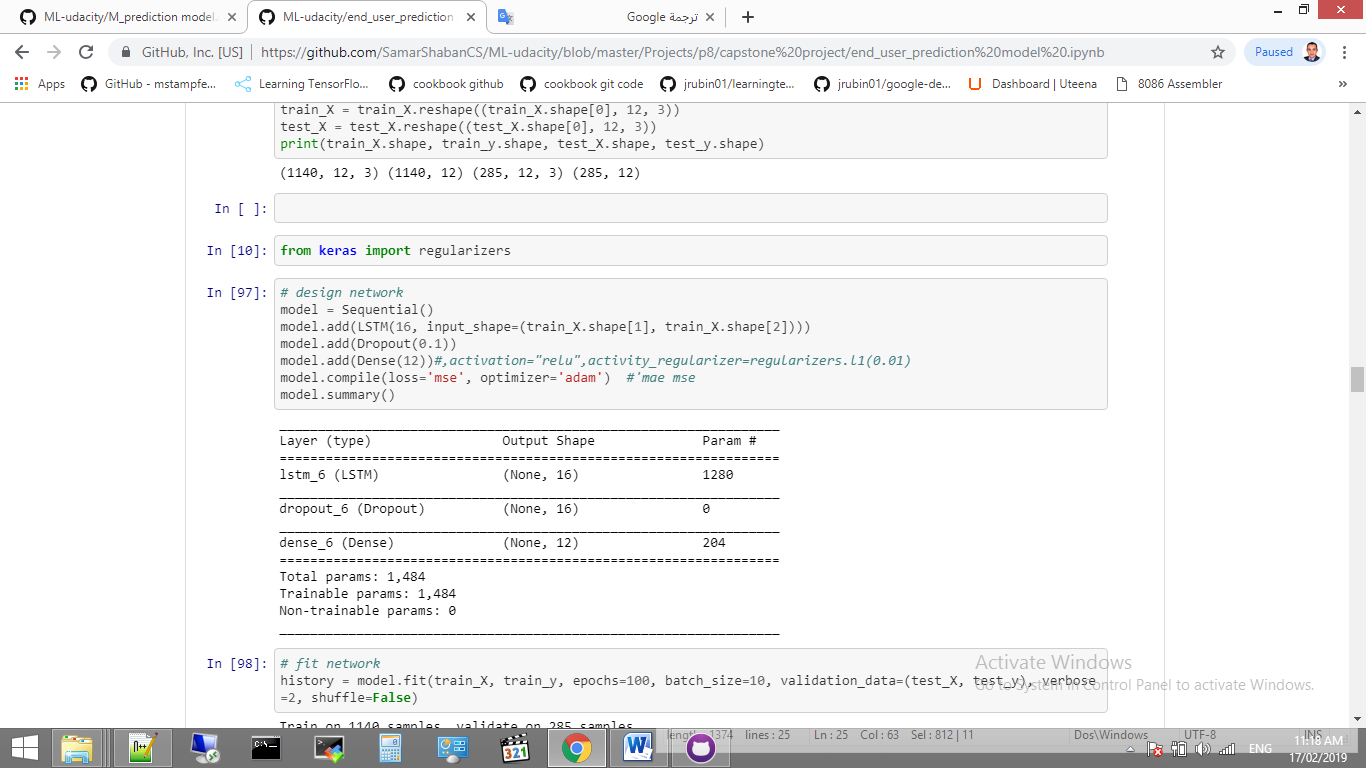
Implementation

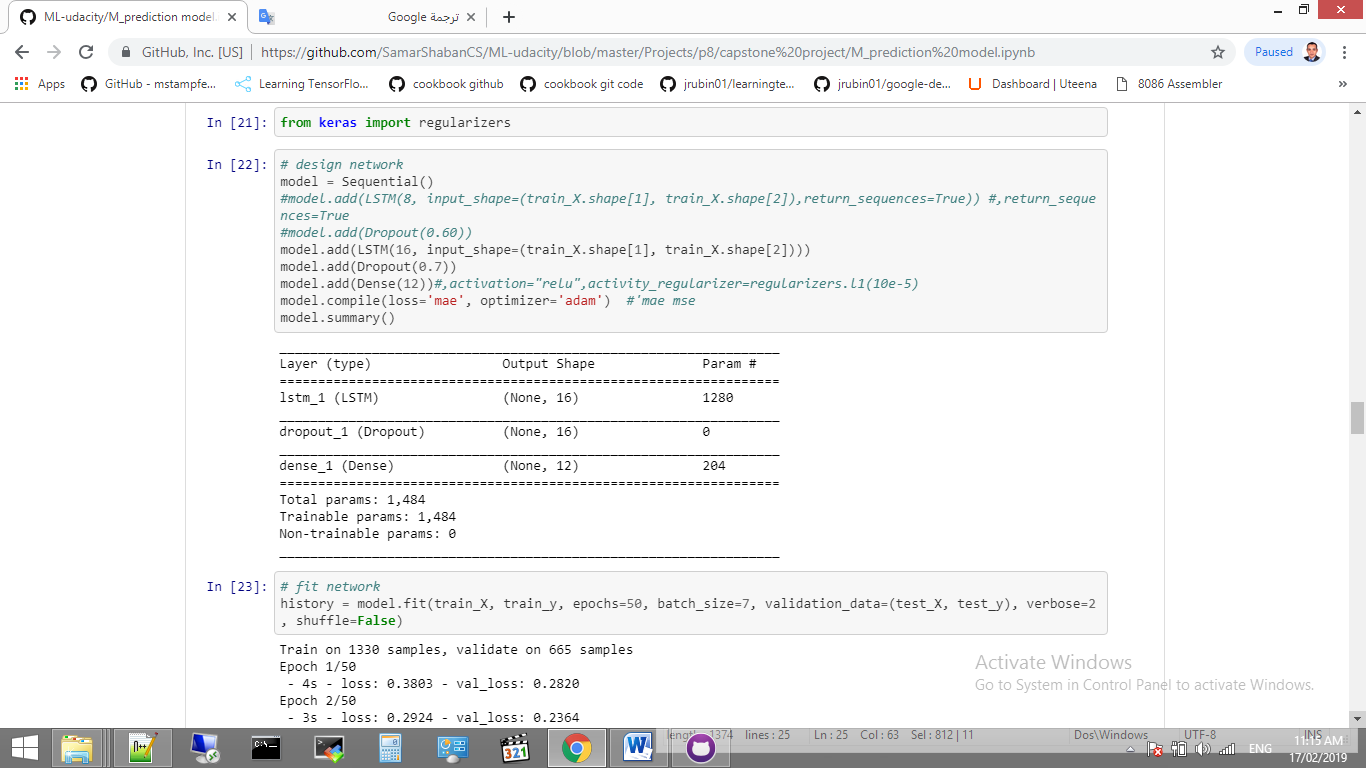
The implementation process can be split into two main stages:

* The predictor training stage
* The application development stage
* During the first stage, the predictor was trained on the preprocessed training data. This was done in a Jupyter notebooks (titled “[M\_prediction model.ipynb](https://github.com/SamarShabanCS/ML-udacity/blob/master/Projects/p8/capstone%20project/M_prediction%20model.ipynb" \o "M_prediction model.ipynb)”: for forecasting the cloud providers’ QoS attributes values and “end\_user\_prediction model .ipynb”: for forecasting the user preferences), and can be further divided into the following steps:

1. Load both the training and validation sequences into memory, preprocessing them as described in the previous section(convert series into (X,Y) or the supervised form, normalize data,..)
2. Define the network architecture and training parameters
3. Define the loss function, MSE
4. Train the network, logging the validation/training loss
5. Plot the logged values
6. If the MSE is not low enough, return to step 2
7. Save and predicted QoS Values

* The following graph is the summary of the network architecture for cloud consumer predictor:



* The following graph is the summary of the network architecture for cloud provider predictor:
* The application development stage which was done in a Jupyter notebooks titled “serviceComposition.ipynb” can be split into the following steps:

1. Load future provisioned QoS attributes values of the 100 providers.
2. Load future preferences QoS attributes values of the user.
3. Split 100 providers services into three classes
4. Define composition function, particle class, particle swarm class (PSO)
5. Run PSO class to obtain the recommended services.
6. Run the brute force algorithm to obtain all possible services composition combination
7. Sort results of brute force algorithm. Assume the top 10 results as the relevant solution
8. Calculate the intersection between PSO algorithm and brute force algorithm
9. Calculate recall and precision according to the following equations

|  |  |  |
| --- | --- | --- |
| **Notations** | Relevant | Non Relevant |
| Retrieved | true positives (tp) | false positives (fp) |
| Not Retrieved | false negatives (fn) | true negatives (tn) |

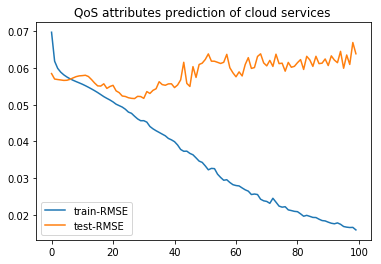
Precision (P) =

Recall (R) =

Refinement

The initial solution has very large RMSE. After adding a drop out layer, the RMSE is significantly reduced.

The following plot of the ​ training​/​validation losses has a divergence indicates overfitting, which can be addressed by adding dropout layer, or reducing the model complexity (e.g. reducing the number of layers), among other techniques.



Results

Model Evaluation and Validation

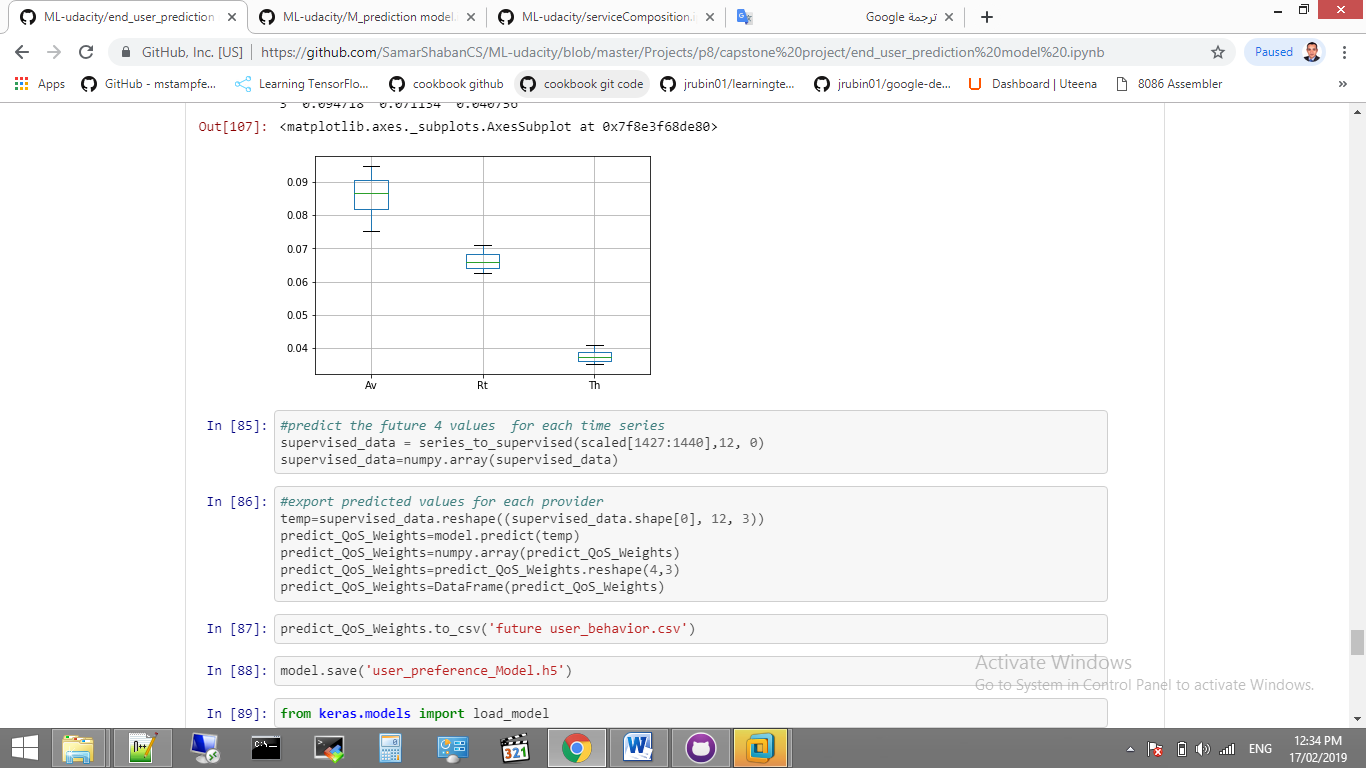
During development, a test set was used to evaluate the model. The final architecture and hyper parameters were chosen because they performed the best among the tried combinations.

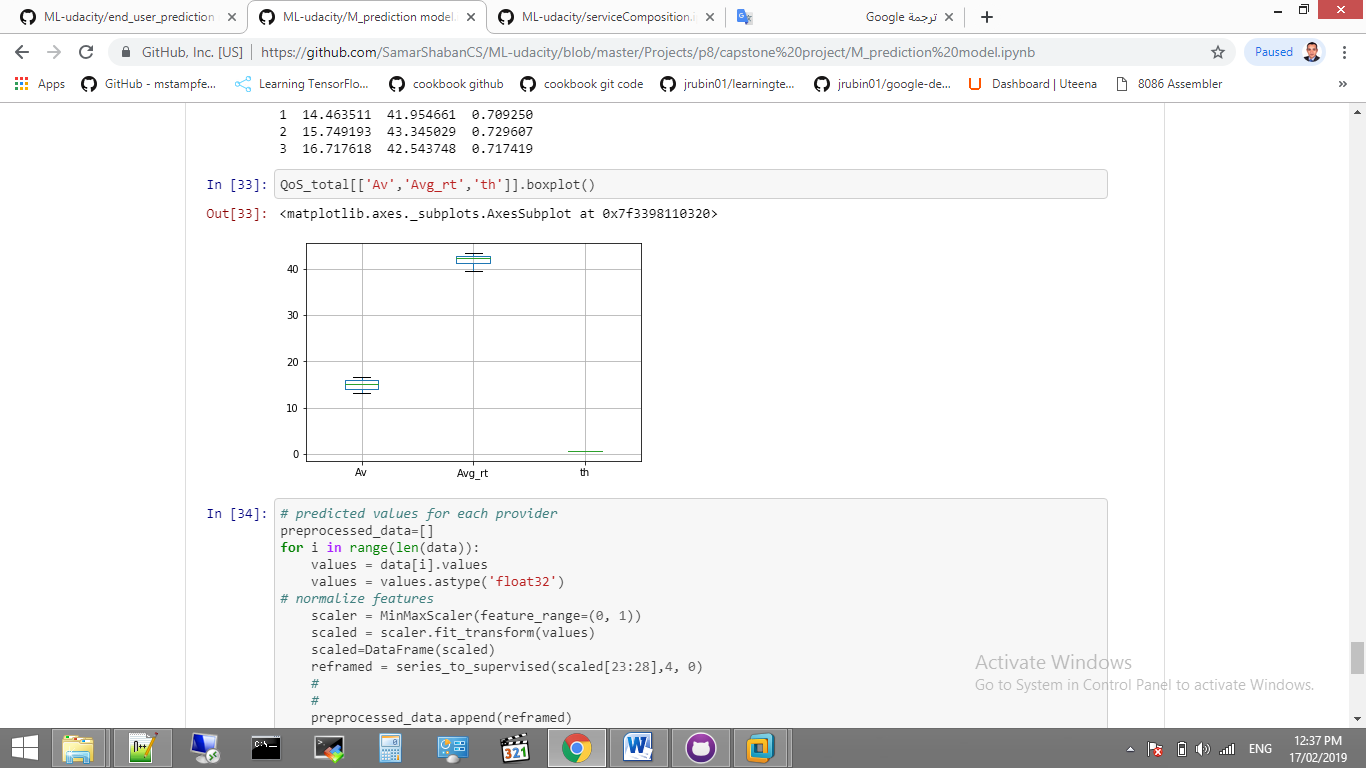
For a complete description of the final model and the training process, the following graph shows the network architecture:

It trained for 50 iterations and each has batch size of 2.

Justification

1. The average test RMSE of the user preferences for all attributes is shown through the following box graph:



1. The average test RMSE of the provider behavior for all attributes is shown through the following box graph:

This is significantly less than the RMSE of the benchmark in all the attributes. The following graphs show the predicted throughput and response time at time t+4.

1. According to the providers provisioned QoS attributes, the average test RMSE of all attributes is shown through the following box graph:

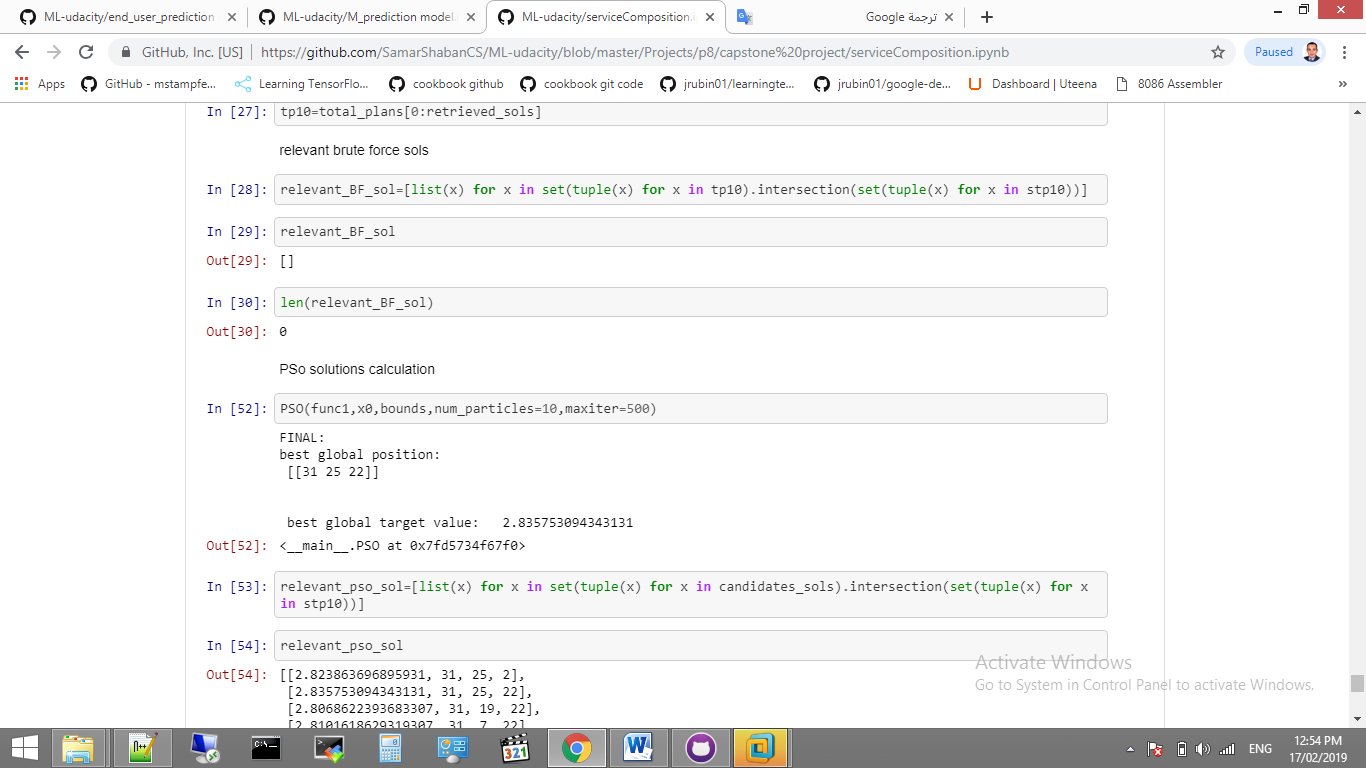
It is seen that the best attributes to use in the composition application is Av, th, and Avg\_rt. They have also less RMSE of the benchmark results.

* The following graph show sample of the availability prediction at time step t+1. It is acceptable results according to the benchmark.
* The following graph show sample of the throughput prediction at time step t+4.

Conclusion

Free-Form Visualization

1. The following graph shows the output of the composition application:



It shows that according to the user preferences: the user should contract

In the first time period with

Provider number 31 at class 1

Provider number 25 at class 2

Provider number 22 at class 3.

Reflection

The process used for this project can be summarized using the following steps:

1. An initial problem and relevant, public datasets were found
2. The data was downloaded and preprocessed
3. A benchmark was determined for the predictor
4. The predictor was trained using the data (multiple times, until a good set of parameters were found)
5. The composition application was adapted to use the results of the predictor
6. Use the composition application to remark the cloud providers and contract with the best ones.

I found steps 4 the most difficult, as I had to familiarize myself with LSTM model which was a model that I was not familiar with before the project.

Also, the data set for both the user and the cloud providers, according to the cloud end user preferences; I use the synthetic generator Timesynth, which is the first time to use it.

According the cloud providers QoS values, the used data set has a small length which is only 28 time slot. This hampers to get high performance (lesser RMSE).

Honestly, I tried to extend the length of the dataset, but I failed.( if you have a practical method that extend the time series dataset to keep the correlation ship between the parameters without using randomization, please tell me)

As for the most interesting aspects of the project, I’m also happy about getting to develop LSTM model using multi variables with multiple lags to predict multiple time step. Also, I’m happy about using Keras which using tensorflow backend, as I believe it will be ​ the​ deep learning library in the future.

Improvement

To achieve high performance we can:

1. Use more capable hardware and use the tensorflow GPU version instead of CPU one.
2. Get more dataset, this will significantly help in getting lesser RMSE.

References:

*Cloud services*. (n.d.). Retrieved August 2018, from , http://www.webopedia.com/TERM/C/cloud\_services.html

Haytamy S.S., K. H. (n.d.). *(2018) ICSD: Integrated Cloud Services Dataset. In: Yang A. et al. (eds) Services – SERVICES 2018. SERVICES 2018. Lecture Notes in Computer Science, vol 10975. Springer, Cham.*

W. Jiang, D. L. (2012). Large-scale longitudinal analysis of soap-based and restful web services. *Proc. Web Services (ICWS), 2012 IEEE 19th International Conference on* (pp. 218–225). Honolulu, HI, USA: the 2012 IEEE 19th International Conference on Web Services.

Z. Ye, S. M. (2016, june). Long-Term QoS-Aware Cloud Service Composition Using Multivariate Time Series Analysis. *IEEE Transactions on Services Computing, 9*(3), pp. 382-393.

1. <https://github.com/SamarShabanCS/Math_for_ML/tree/master/time%20series%20data%20QoS> [↑](#footnote-ref-1)
2. <https://ieeexplore.ieee.org/document/6964807> [↑](#footnote-ref-2)
3. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> [↑](#footnote-ref-3)