Assignment DSESK-101

Title: Implementation of regression techniques on time-series data to generate future predictions.

Consider the following hypothetical discussion.

A computational system records transactions. These transactions are stored in a time series log. Now consider following scenarios,

- Scenario 1: In a grocery store, a sales purchase is made for some goods. This transaction with relevant information is recorded in a log by a computer system. This log then contains historical information of all the goods purchased in that store.
- Scenario 2: A Cloud platform continuously records the resource utilization of instances per minute and stores this information in the resource utilization log. For example, Instance A, has its CPU, Memory, Disk and Network Bandwidth utilization with a timestamp stored in a log file. This log file then contains the historical data of resource utilization from the time that Instance A was created and till the time it was destroyed.

Scenario 1 and Scenario 2 are completely different, and are only related by the fact that forecasting is a useful tool to conduct strategic and optimization planning. Assume relevant variables to understand the problem statement, for example, in scenario 1, overall sales forecasting would require transactional logs of all the sales, and cosmetic sales forecasting would require logs related to cosmetics. Where as in case of scenario 2, predicting future resource utilization for a single instance would require the resource utilization of that single (here) instance A, where as to predict resource utilization of a project/group, resource utilization logs of all the instances in that group (say) instance A, B, C, D, E, F will be required.

Problem Statement:

Given scenario 1 and 2, How can we use Machine Learning constructs, to implement an intelligent system, that allow the user to see a predicted forecast of transaction. Input to the system is transactional log files. The system is expected to perform required data cleaning and transformation. Then the system is expected to perform required analysis on the processed data and generate results in terms of future predictions. The output of the system should be a predicted transactional log.

Functional Requirements

1. System should convert given data into required data model. This stage generates `input-data`.
2. System should implement relevant and required data cleaning and data transformations techniques in the input data. This stage generates `pre-processed-data`.
3. System should (if required) perform training on pre-processed-data and generate machine learning model. This stage will generate a `model`.
4. System should process pre-processed-data and generate prediction results. This stage generates `predicted-transactions`.

Expected Results

Given the data-set below, following are the expected results from the system,

1. Predictions per Instance (required)
2. Predictions per Group (required)
3. Predictions per Resource (CPU/Memory/Network/Storage) per group or per instance (optional)

Expected Deliverables

1. Link to the code base as a GitHub/GitLab repo. (required)
2. Assignment documentation which should include the following, (required)
    a. Problem Statement
    b. Brief description (your understanding about the problem and its brief explanation)
    c. Proposed solution (this section must include your solution to the problem, relevant algorithm, techniques, frameworks used)
    d. Block diagram (a diagram explaining, data flow, processing and overall architecture of your solution)
    e. Sample Test Results
    f. Future scope (possible improvements, performance tweaks, ability to process other time-series data etc)
3. Document containing references, bibliography items. (optional)

General Expectations

1. It is expected to implement this assignment (partially or completely) in Python programming language
2. Any information that is not available in this document can be assumed, and explicitly mentioned in assignment documentation

Data-set

The data-set to be used for this assignment can be downloaded from here

This hypothetical data-set represents transactional logs of resource utilization of over 3000 instances. This data-set includes over 3000 folders, wherein each folder has a unique name, e.g. `group_4_506294bf-c2d1-4c3e-887f-ab10f04908d7` where `group_4` represent the group ID and instance `506294bf-c2d1-4c3e-887f-ab10f04908d7` belongs to that group ID. This folder has a file named as `mem.log` which contains resource utilization records. Each record is of the format,

```
"{timestamp}:{Memory Allocated}:{Memory Used}:{CPU Allocated}:{CPU Used}:{Network bandwidth utilization}:{Storage space utilization}"
```