STAT995 – Batch Machine Learning

Reflective REPORT

Student Name: Minnie Cherry Chua Tan id:17996397

Supervisor: Victor Miranda 13 june 2022

2022

Contents

[I. Introduction 1](#_Toc105781544)

[II. Project Evaluation 2](#_Toc105781545)

[A. Evaluation of Project Against Proposal 2](#_Toc105781546)

[B. Significance of the Project to the Organization 2](#_Toc105781547)

[C. Critique of the Methodology 3](#_Toc105781548)

[III. Personal and professional development 3](#_Toc105781549)

[A. Technical 3](#_Toc105781550)

[B. Social 4](#_Toc105781551)

[C. Economic 5](#_Toc105781552)

[IV. Conclusion 6](#_Toc105781553)

[V. Appendix 6](#_Toc105781554)

[A. Proposed and Actual Timelines 6](#_Toc105781555)

[VI. References 8](#_Toc105781556)

# Introduction

The project is about batch machine learning that processes one record at a time using various models: regression, classifier, rules, samples, text, neural network, unsupervised learnings, which I developed from scratch based on accumulated learnings at AUT, and application of online learnings of various scripts to create an OOP out of those scripts to create the tool. The success of the project can be attributed to the course undertaken in AUT to provide me knowledge in maths, AI and its metrices used for analysis with usage of different tools (Weka, Matlab, Saas, R, Python, Neucube), knowledge sharing online through Google, additionally online courses I’ve brought in Datacamp, PyImageSearch, Udemy, studied and I applied in batch OOP machine learnings development (*Machine Learning A-Z™: Hands-On Python & R In Data Science* for scikit-learn, *Deep Learning A-Z™: Hands-On Artificial Neural Networks* for ANN, *Deep Learning Convolutional Neural Networks in Python* for CNN, *Deep Learning Recurrent Neural Networks in Python* for RNN) The technique used is to use existing scripts examples to be made generic using implementation of interfaces in OOP to have the skeleton of the examples. After the paper would be discussing about the project evaluation, personal and professional development in undertaking this project.

AUT as an educational institution has many experts in their own fields. There are many tools available that may do the training and testing of the models, which is either a GUI based, or in script based, one or many. The batch machine learning has its usability that it can be dynamically setup to perform multiple train-test of various algorithms and/or methodologies at one time.

The aim of this project to ease development process and make routinary tasks automated by configuration setup of input files with customization if needed, instead of doing multiple scripts development alone:

1. To minimize user interaction and/or development by creating a usable tool for user to add or update only input files to execute instead of scripts, that produced graphs and excel metrices with minimal customization or development required
2. To have saveable retrievable results organized and archived by name, date, time
3. To have the statistical metrics in selection of different modelling techniques used
4. To use hyper-tuning using GridSearchCV for some scikit-learn samples whenever applicable, and/or elbow method to get the optimal cluster for the clustering. Also, tensorflow don’t have automated hyper-tuning
5. To provide software tool usage with some sample input files as templates, so user can regenerate the results generated in the document via the developed tool.
6. To do some deep learnings like ANN, CNN, RNN to handle image, text processing and other unsupervised techniques like SOM, RBM.

# Project Evaluation

The project evaluation would be based on: (1) project achievements against those outlined in the original project proposal, (2) significance of the project to the organization, (3) critique of the methodology applied in the project.

## Evaluation of Project Against Proposal

There is some diversion from the proposal and actual targets, and its timeline, but it’s to improve and update the batch machine learning to have more features and make them more robust as per table below. Some major improvements are that deep learnings are taken into account, neural network in tensorflow, unsupervised learnings in SOM, RBM. Though communication with one week lead time would have been better with any changes that can be attributed to the fact of the time to learn, and do it as well. Also, the unforeseen environment issues are never accounted for. The baseline of the proposal is met overall having a batch machine learning with example templates of (1) *regression* - LinearRegression, PolynomialFeatures, SVR, DecisionTreeRegressor, RandomForestRegressor, (2) *classifier* - LogisticRegression, KNeighborsClassifier, SVC, GaussianNB, DecisionTreeClassifier, RandomForestClassifier, (3) *cluster* - DBSCAN, AgglomerativeClustering, KMeans, MiniBatchKMeans, (4) *rules* - apriori, eclat, (5) *sample* – UCB, ThompsomSampling, and deep learnings of: (6) *text* (natural language processing) – using classifications of GaussianNB, RandomForestClassifier, (7) *neural network* – ANN, CNN, RNN (SimpleRNN, GRU, LSTM), (8) *unsupervised learnings* – SOM, RBM.

## Significance of the Project to the Organization

The significance of the project to the organization is to have the power of dynamic setup of the data and parameters both values and functions to minimize the routinary task of developing multiple scripts. As per the objectives:

|  |  |  |
| --- | --- | --- |
| **Objectives** | **Met** | **Comments** |
| - setup, with customization as needed | ✓ | For the models and examples done, as to the limitations of time without all the knowledge |
| - saveable result | ✓ | Automatically saves the excel and graph when setup, based on filename, date and time with the filename as the folder name for individual models, “file\_stats” for the global stats whenever applicable |
| - statistical metrics | ✓ | Spool the metrices whenever applicable |
| - hypertuning | ✓ | Done for scikit-learn whenever applicable, i.e. cluster, didn’t apply for tensorflow as cannot be applied |
| - provide input templates | ✓ | Provided for the example data modelling, which can be used as a template with similar file setup, so data files can be changed accordingly |
| - apply deep learnings | ✓ |

Through the limitations of resources like time, the tool is developed to the acquired incremental knowledge over time that is limited with the development and testing of the modelling and examples. So, more modelling, data needs to be tested to gain the likelihood of the tool. There’s power in data to be seen as homogenous and heterogenous to be grouped together as one can do and repeat the same process with the other datasets as long as it’s dynamically setup, then, development of routinary tasks or having multiple scripts can be minimized, which can be efficient and effective in the long-run. It can also be used as a reference of techniques applied in the development, and different modelling used, or reuse, recycle the templates.

## Critique of the Methodology

Abstraction Layered Architecture is used to avoid and minimized tight coupling using 4 layers of application layer, domain abstraction layer, programming paradigm, programming language (libraries). (Spray, n.d.) The initial plan is to do a MVC (Model, View, Controller) out of the batch machine learning. Due to the fact, there’s a conflict of time and design for input text file based as originally done. Only implementation of the interfaces are used, coupling it with JSON, dictionary, pipeline, dataframe, regex manipulation, eval, dir() and \* to make a generic and dynamic parser of functions and values in the input files. At the same time, adding another layer of UI view, would addon to another level of complexity, which can limit the dynamic mechanism of the tool, since it needs to be mapped for the functions and parameters. What if the functions and/or parameters are decommissioned? Then, the setup of mapping the functions with its parameters in the UI may become invalidated unless its maintained or just per version of the libraries used.

Due to time-constraint, only specified modelling methodology and examples are developed and tested in the batch machine learning as specified. However, with the usage of \*, the setup can be made dynamic in the input files as long as its part of the package module as well as the usage of regular expressions. So, for examples, layers can be added in the neural network with the similar files. It’s not totally zero-coding, as with different datasets and file types, preprocessing of those files maybe needed to be customized. But, once customized, it can work for the similar setup without development. Unless limitations are put in the tool to have certain file constraint or input to be processed.

# Personal and professional development

## Technical

|  |  |
| --- | --- |
| Concrete Experience | To think in a box diagram design before development helps to ease the development flow. To implement the batch machine learning in interfaces helps isolate dependent and independent functionalities of each group or category via input files, so additional methodology can be added easily. Though the dependent methods that are shared between modules need to be retested even for the previous working ones as unit testing are not in-placed. Though logging mechanisms are placed in the tool, it’s not fully optimized as it can accumulate over time, so print and exit are used often during the development and testing cycle for efficiency and effectiveness, and quicker turnaround. |
| Reflective Observations | Software architecture design is an important initial step before starting with development for it sets the structure and limitations. During the development of the tool, design is done, but, only in object level, not in the attribute level, which can be useful in making things more standard and uniform. The design level can still be improved, though it’s working and usable. As Ellen Ullman said, ‘*We build our computers the way we build our cities—over time, without a plan, on top of ruins*.’ (goodreads, 2022) Top-down approach being configurable in setup is done to pre-process, build the model, train the model, predict the model, provide necessary outputs. To provide unit-test and more informative logs can be more useful, but constraint in time to build more input templates to meet the requirement of the proposed batch machine learnings. |
| Abstract Conceptualisation | Environment issues are encountered time and again that my laptop needs to be reset to factory settings several times. To place it in a cloud environment and version control would be nice to have, which should have been considered at the start of the project intake. It wasn’t done for this project since it was an individual project and the technologies I’ve used before professionally are legacy, not in cloud environment and not in git, except for personal technical exploration before I used a bit of cloud, and git. |
| Active Experimentation | To have active experimentation is alright during the development cycle to know which way works best efficiently and effectively even in future projects, that are done in some instances of the development. But, to have a predefined scope and limitation would help in consuming valuable time and effort. |

## Social

|  |  |
| --- | --- |
| Concrete Experience | Since the project is an individual project, the social aspect is very limited. One aspect to socialize is to communicate with my adviser of the project progress, which is done usually weekly, if not, at times fortnightly that is normally done through emails. On occasions, face to face meetings is also scheduled ahead of time. |
| Reflective Observations | To have online meetings like the mid-project review and online communications for the project progress works well for me as I’m introvert, so speaking can be an extra effort for me or I can forget what to say as I at times get nervous with people around. At the same time, to have the update in writing gives the client and myself, the traceable history of the progress as its recorded. Though sometimes to have the communication done verbally can be a lot efficient and effective as one will have the answer and feedback immediately compared to email communication. So, on occasions face to face meetings are conducted. |
| Abstract Conceptualisation | To be clear in the email must be done, as with the weekly project progress for the purpose of not confusing my adviser, since he’s a professor, on top of the other things he does so he’s a busy person. Later, on few occasions, when there’s only document update, or just an input files, I give multiple updates (2x, 3x) for the document alone in a week, which I realized I should have done one document update for a week, with a clear statement of what have been done. Also, for the project instead of giving him a zip file, it’s better to have done in a cloud environment, to replicate the results of the tool. Not only from my end, but from his end as well, which is overlook as it’s an individual project. By having the environment, library module setup can be minimized, though pipenv, pip lock files are provided later. As advised, I need to be clear when it’s already a final update. |
| Active Experimentation | To know what went well and what didn’t with the reasons accordingly, the lessons learned can be applied for any future work or study. Lessons learned: (1) Online communication like emails or other forums – must be *clear* and *concise* with a template preferably and a common environment to share whenever applicable. (2) Real-time communication must be *prepared* not only for the materials used, but for any technical glitches or any issues that may occur, (3) Communication is two way, which most amiss that it needs to have *feedback*. |

## Economic

|  |  |
| --- | --- |
| Concrete Experience | Person with good technical skills who continually improves are hard to find. At times, such a person can do few people’s jobs. With the exponential growth in internet usage, so does data consumption increase exponentially. By 2025, data generated each day is expected to reach 463 exabytes globally. (Vuleta, 2021) Google, Amazon, Facebook store at least 1200 petabytes of information and the world spends almost $1 million per minute on commodities on the internet. (Vuleta, 2021) AI via machine learnings helps make the raw data into useful information to support decision making of the executive levels and management. Moreover, useful information can bring in knowledge discovery and/or rediscovery and/or information confirmation. Data makes the world evolve, so, making a good use out of them is pertinently important. In New Zealand, data analyst earns about $80K-$130K per year, and data scientist earns $100K-$160K per year, whether it’s healthcare, education, manufacturing. (careers, 2022) |
| Reflective Observations | People in school are usually scholars or have doctorate degrees, or highly intellectual so they can wear multiple hats, do more than one person’s work. Given the fact that AI market is expected to grow by 60B by 2025, by 2030, global GDP will grow by $15.7T due to AI (G., 2022), the demand for AI skills will also increase, especially with the tough market competition in the industry. So, the industry offer can be more attractive economically than inside the institution, except if the person’s perspective is more than for the economic benefits. But realistically, it can be a tough competition. If one skilled person is lost in an institution, the lost can be more than one’s person’s jobs. So, to have a tool like batch machine learning as a tool can be useful. |
| Abstract Conceptualisation | The following learnings can be applied for future endeavours. (1) To know the pay scale of AI/ML person is important to know what and how to demand. (2) To know the increasing need of those jobs would make one desires to upskill more in AI/ML. As per World Economic Forum, AI will create more jobs than destroy it, by 2022, 75 million jobs are displaced by automation with 133 million emerging ones. (Brown, 2019) (3) To be able to retain people with skills can be productive and beneficial to any organization. (4) To know the industry outlook overall says that the market demand for such skills would grow exponentially. So, overall, to be able to use a tool of batch machine learning that can be dynamically setup can be beneficial, economically among others as to having many experts. |
| Active Experimentation | In future or any subsequent project, to make the project shareable in source control storage with better application of the software architectural design are preferrable. Because no one would like to create and develop a tool that no one will use. Moreover, the value never diminishes when things are done right. |

To know alone is not enough: (1) what went well, (2) what went bad, and (3) the reasons behind them. But, to be able to maintain and/or improve the good, minimize or eliminate the bad are what’s the knowledge is called for. Like in the business term, Kaizen (改善), renew for good or continuous improvements. (Kakihara, 2021) It’s in the act of humility that improvements occur, not in a leap of improvements, but, little acts of improvements that it become into the milestones of improvements.

# Conclusion

Currently, the batch machine learning can process 8 types of modelling in a batch, with about 45 examples in various algorithms in about 30 minutes to 1.25 hours using around 2-4G free memory, which is in a way effective and efficient as there are some deep learnings in the sample modeling. Nevertheless, it may be optimitized further.

Doing the project, though it’s an individual work, won’t be possible if the opportunity is not given by my adviser, and for the incremental accumulated learnings over time. The batch machine learning is working. Nevertheless, in any development, it can never be claimed it’s 100% perfect. Doing the project is not about my performance in the project, but also the experience that comes with it, both good and bad. On how that can improve me as a person both individually and professionally in the aspect of technical, social and economic classified by Kolb Learning Cycle of Concrete Experience, Reflective Observations, Abstract Conceptualisation, Active Experimentation. For if one can’t view on both sides, the good and bad, how, then, can one improve?

Even though, it’s an individual project, never take all the credit by yourself. Because the reality of life, we are all sitting on top of the knowledge of someone’s else, whether directly or indirectly. As I may define it, we are like ants in top of an elephant. Kaizen is a good mantra in life. Also, in any project, even in individual project, one needs to connect to someone in some way possible to make a project feasible, else, it’s just self-learning.

Final words, kaizen and no one is one’s own.

# Appendix

## Proposed and Actual Timelines

|  |  |  |
| --- | --- | --- |
| **Week** | **Original List** | **Updated List** |
| Week 1 |  | Updated the version 2 the base version created on the prior dates from the version1 development done – created the static pipeline’, debugging |
| Week 2 | Proposal documentation | Proposal documentation, updated to have the dynamic parser with different implementation’, debugging |
| Week 3 | Add/Update document, main, config.json, input.txt | Add/Update document, updated the preprocessor module for the missing data to be dynamic execution’, debugging |
| Week 4 | Add/Update document, input.txt – test different types of files, helper, constant | Add/Update document, updated the preprocessor execution to have configuration for high ordinality values via GLMMEncoder(), scaling, added static feature selection and hypertuning, added Feature graph, debugging |
| Week 5 | Add/Update document, preprocessor, helper, constant, graphics | Add/Update document, preprocessor: updated feature selection PCA, SelectKBest to have the scores in excel and savable graph of the scores, updated the hypertuning to be dynamic, combine the regression and classification’, debugging |
| Week 6 | Add/Update document, preprocessor, scale, helper, constant, graphics | Add/Update document, preprocessor: hypertuning of dynamic scores, spool in excel, reading of non-MIME datasets of fetch (real-world) and load (sample), use measure\_performance as a decorator’, debugging |
| Week 7 | Add/Update document, preprocessor, feature, helper, constant, graphics | Add/Update document, updated model interface to have metrices, added regressor – LinearRegressor, updated main() to have classification types between Regressor and Classifier’, debugging |
| Week 8 | Add/Update document, regression, helper, constant, graphics | Add/Update document, added regressor to have PolynomialFeatures, SVR, DecisionTreeRegressor, RandomForestRegressor, smoothing for the graph applied only for SVR, DecisionTreeRegressor, RandomForestRegressor, batch testing, separated metrics of regression and classification’, debugging |
| Week 9 | Add/Update document, regression, helper, constant, graphics | Add/Update document, added classifier to have LogisticRegression, KNeighborsClassifier, plot for contour and scatter, plot report for confusion matrix, classification report, prediction error, roc\_auc’, debugging |
| Week 10 | Add/Update document, classifier, helper, constant, graphics | Add/Update document, added classifier to have SVC, GaussianNB, DecisionTreeClassifier, RandomForestClassifier, plot for kde with scatter’, debugging |
| Week 11 | Add/Update document, classifier, helper, constant, graphics | Add/Update document, added cluster: KMeans, placed in the title ‘scatter plot’, debugging |
| Week 12 | Add/Update document, cluster, helper, constant, graphics | Add/Update document, cluster: Agglomerative, debugging, enhance to point to different input files, ppt presentation, added plot\_dendograph() |
| Week 13 | Add/Update document, cluster, helper, constant, graphics | Add/Update document, develop, testing, debugging cluster: MiniBatchKMeans, DBSCAN |
| Week 14 | Add/Update document, cluster, helper, constant, graphics | - |
| Week 15 | Add/Update document, rules, helper, constant, graphics | Add/Update document, develop, testing, debugging rules: APRIORI, ECLAT |
| Week 16 | Add/Update document, rules, helper, constant, graphics | Add/Update document, develop, testing, debugging sampling:  UCB, Thompson Sampling |
| Week 17 | Add/Update document, neural network, helper, constant, graphics | Add/Update document, develop, testing, debugging natural language processing:  Porter Stemmer |
| Week  Summer | Add/Update document, neural network, helper, constant, graphics | Add/Update document, develop, testing, debugging neural network: ANN |
| Week Summer  /  Week 18 | Add/Update document, neural network, helper, constant, graphics | Add/Update document, develop, testing, debugging neural network: ANN neural network – update graph to have validate\_data,confusion matrix for discrete, plot3d for continuous number, added CNN (study and implement as I do)  Updated other files: helper, constant, graphics |
| Week 19 | Add/Update document, neural network, lstm , helper, constant, graphics | Add/Update document, develop, testing, debugging neural network: loading with language encoding, confusion matrix, CNN for natural language processing, and loading to local directory, retest sample tests, change config.json to have base directory |
| Week  20 | Add/Update document, neural network, lstm , helper, constant, graphics | Add/Update document, develop, testing, debugging neural network: to include y-test and y-prediction in XLSX tab if applicable, updated CNN for one sample based on filename if no y in the XLSX, study RNN – LSTM, added load\_files(), preprocess\_target() |
| Week 21 | Add/Update document, neural network, lstm , helper, constant, graphics | Add/Update document, develop, testing, debugging neural network: to update RNN – LSTM: updated preprocessor.py: preprocess\_target(), added to\_dataframe(), neural.py: added plot\_target() and LSTM handling |
| Week 22 | Add/Update document, neural network, lstm, helper, constant, graphics | Backup, Reset / format PC for 2 days as reset twice still getting high memory usage , Add/Update document, develop, testing, debugging regression\_classification.py as the ax.plot in subplot becomes unavailable so use seaborn instead in my PC by switch server; and debugging issues in regression (regression\_classification.py) for 2 days; review SOM, added nonsupervised.py for SOM handling today (1 day) |
| Week 23 | Add/Update document, test for dataset of assignment1 heart, mortgage | Add/Update document, develop, test for SOM to include summary\_stats, plot of confusion matrix, scatter |
| Week 24 | Add/Update document, test for dataset of assignment1 heart, mortgage | Study RBM, Auto-encoder, Add/Update document, develop test for RBM in the preprocessor, adding input\_file, added the initial development for RBM using Gibb Sampling in nonsupervised.py – added some functions convert(), evaluate(), rate\_data() |
| Week 25 | Add/Update document, test for dataset of assignment2 question 1 | Add/Update document, develop, test for RBM – nosupervised.py - Added plot\_rbm(), updated evaluate() - add history of loss and accuracy score to xls, updated rate\_data() to make the scoring setup generic; model.py - updated plot\_cm() to check for if cm\_df exists to prevent errors |
| Week 26 | Add/Update document, test for dataset of assignment2 question 1 | Add/Update document, develop, test further for RNN – updated model.py, preprocessor.py - load\_auto\_generated\_by(), constant.py, neural.py - multistep\_forecast() |
| Week 27 | Add/Update document, test for dataset of assignment2 question 2 | Add/Update document, develop, test further for RNN – updated model.py, preprocessor.py, neural.py - to handle to RNN, GRU, LSTM; created powerpoint, pipenv for project virtual environment |
| Week 28 | Add/Update document, test for dataset of assignment2 question 2 | Checking on video making, having some software issues, i.e. speaker, microphone in my laptop; requested computer shop to revert to windows 10 |
| Week 29 | Add/Update document, test for dataset of assignment2 question 2 | Updated the filename to ANN if the built model has Dense only, updated multistep\_forecast() for ANN |
| Week 30 | Add/Update document, test for dataset of assignment2 question 2 | Update plot\_cm() to put label text for specific example, Updated plot\_confusion\_example() to add the index in the title, updated powerpoint, updated samples to add xlsx, updated document, added ads\_selected\_count() |
| Week 31 | Add/Update document, test for different dataset | Updated project document – abstract, acknowledgment, result analysis, com; input file of sample; created the reflective report document |
| Week 32 | Add/Update document, test for different dataset | Updated the reflective report document, cleanup unused codes |

# References

Brown, T. (2019, October 11). *The AI Skills Shortage*. https://itchronicles.com/artificial-intelligence/the-ai-skills-shortage/

careers. (2022, April 8). *Data Analyst / Scientist Earnings*. https://www.careers.govt.nz/jobs-database/it-and-telecommunications/information-technology/data-analyst

G., N. (2022, June 2). *101 Artificial Intelligence Statistics [Updated for 2022]*. https://techjury.net/blog/ai-statistics/

Kakihara, S. (2021, August 2). *What Is Kaizen in Japanese, and How Is It Useful?* https://thetruejapan.com/what-is-kaizen-in-japanese-and-how-is-it-useful/

Spray, J. (n.d.). Abstraction Layered Architecture:Engineering the structure of software. *Abstraction Layered Architecture*. http://abstractionlayeredarchitecture.com/#truelayers-instead-of-hierarchical-decomposition

Vuleta, B. (2021, October 28). *How Much Data Is Created Every Day? [27 Staggering Stats]*. https://seedscientific.com/how-much-data-is-created-every-day/