DSA4299 Internship Report
Micron – Global Procurement Data Analyst

ShaoMin Liu
A0183798E

April 24, 2021

Contents

1 Abstract

I am very lucky to have the opportunity to work in Micron, Global Procurement Data & Analytics(GPDNA) in the past 5 months as intern. It was a fruitful experience that allowed me to see data analytic in business application. During the internship, there were many opportunities for me to apply what I have learned from NUS and there were also challenges that required learning new things from scratch.

This report attempt to summarize the internship experience that I had in Micron, present the projects that I had completed and finally, elaborate on the learning points that I have taken away from this journey.

2 Introduction

2.1 Micron, Global Procurement Data & Analytics (GPDNA)

Micron is an industry-leading semiconductor manufacturer that delivers memory and storage solutions all around the world. Micron has two Fabs in Singapore that employs approximately 10,000 staffs.

The Global Procurement team is responsible for the purchases and spending of the firm in all aspects. From raw materials in producing the chips to the tables and chairs we use in the offices, Global Procurement will source for suppliers, negotiate for prices, coordinate the shipments, payment for goods and many other responsibility. Global Procurement play a critical role in the company's worldwide operations.

As part of Global Procurement, GPDNA is a team dedicated to providing analytics support to the department, as well as maintaining data security and data governance. GPDNA produces reports and visualization using Power Apps and Tableau to help Global Procurement department in her decision making processes. In additional, GPDNA also build and maintain ETL(extract, transform, load) pipelines to provide data to stakeholders.

2.2 Intern Responsibilities

As an GPDNA intern, my responsibility was to develop efficient solutions for Request Tickets to support the business operations in Global Procurement. In the development of the solutions, I had to make sure that my solutions were scalable and expandable in the future.

Besides my day to day business tasks, I was also responsible for experimenting and prototyping some new projects that will improve the team's efficiency or bring business value to Global Procurement.

3 Projects

In the past 5 months, there were a few request tickets assigned to me. In this section, I will focus on only three of the major projects that has taught me the most amount of things in the process.

In the explanation of the deliverables, I will first explain problem that the project was trying to address. Then I will elaborate on the how I have came to arrive at the solutions. Last but not least, the outcome of the solution.

3.1 Airflow Infrastructure

3.1.1 Business Opportunity

GPDNA maintains a set of ETL tasks that is growing gradually in complexity and quantity. The original solution was using an internally maintained script to trigger the job automatically using Cron.

Since it was internally maintained by GPDNA, we enjoyed the full extent of flexibility to tailor make Cron Job according to demand. The pitfall was that it became increasing challenging to maintain each Cron Job as there are more and more of them. The nature of the Cron tool also dictates that it is unfriendly to new team members to pick up this tool.

3.1.2 Airflow

Therefore we decided to use the open-sourced tool developed and maintained by Airbnb — Airflow, to manage our ETL tasks.

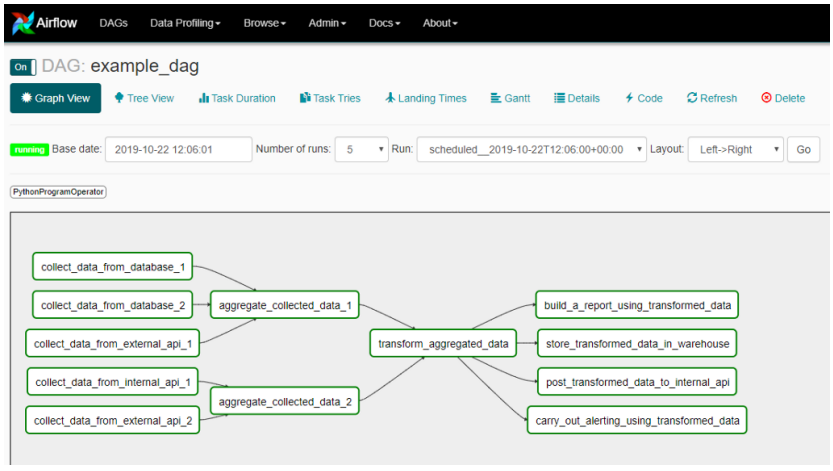


Figure 1: Typical ETL process

Airflow is a tool that runs on Python, it has a user-friendly UI that let us monitor the status of each task in real-time. Using Python, we can also design complex ETL Job comprises of different tasks. Because Airflow is professionally maintained by Airbnb, it supports many useful functions such as email warning during a service failure, or ETL job report during a breakdown to identify the error point.

3.1.3 Setting Up Airflow

In order to isolate development environment from the production environment, it was ideal to set-up Airflow using Docker containers. This was proven to be much more complicated and difficult than I initially imagined. Both Airflow and Docker are tools that were completely foreign to me, so I spent more than a week studying the tools and experimenting them.

After setting up a simple Airflow container on Docker, I was able to run some scripts using the container already. But there was a problem that some of the tasks that were designed to run in parallel were not running concurrently. After some investigation, Airflow uses a **Sequential Executor** by default, of which, runs all given task in sequential order. It was clearly not the most efficient way of running the tasks. After doing some research, I realized that I need to use a **Celery Executor** to allow parallel execution on the machine. But in order to set-up a **Celery Executor**, a Postgres container is needed for back-end database. Therefore I will have to learn to set-up a Postgres container to host the database service for the Airflow.

The hardest part came after the main set-up was completed. The ETL scripts that we currently comes with their individual dependencies. Onboarding the dependencies was a huge challenge because some of the dependencies were no longer supported and it was not so user-friendly to install them on Docker. An example will be the PyOdbc package that was needed to communicate with Microsoft SQL Server. PyOdbc requires a driver to connect to the SQL Server, most of the drivers were designed for Windows platform. But the docker image that I used had a light-weight Linux core, therefore I had to learn how to compile the driver from its source code so that it is usable on the docker image.

3.1.4 Result

Using Airflow, we have a very user-friendly interface to manage our active and past performance of ETL. It is very easy to track when are the tasks triggered and how much time do they take to finish the task. The logging feature also allow us to pinpoint the bug when the script is failing. Airflow can also send a email notification when a script is failing or automatically re-try the tasks if they fail. Airflow also has many Sensor object implemented so that tasks can be triggered based on non-time conditions. This remove the need for us to

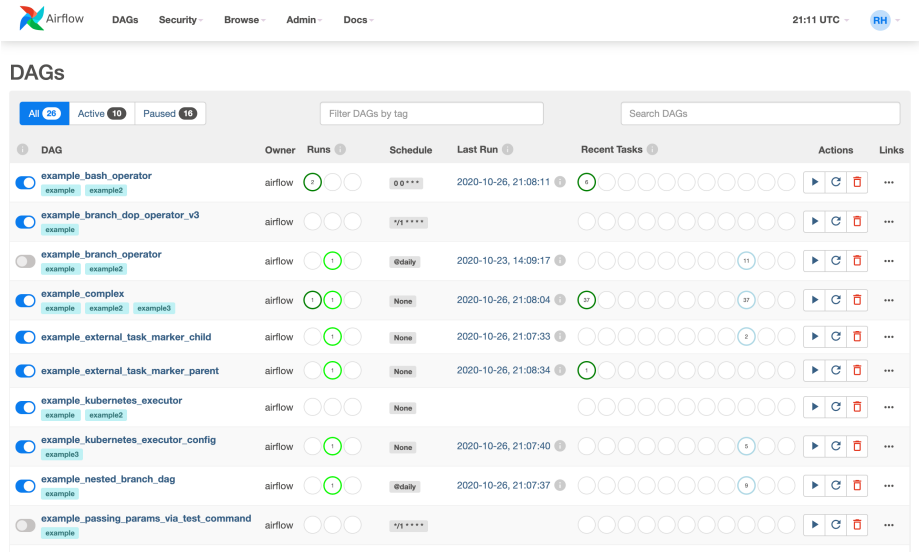


Figure 2: Airflow UI Page

re-invent the wheel to implement these features and focus on our main business problems.

As of the day I am writing this report, GPDNA has two ETL pipelines that are currently running on Airflow, more ETL pipelines will be gradually migrated to Airflow in the future.

3.2 Weeks of Supply Forecast Report

3.2.1 Business Opportunity

In procurement, controlling the amount of material purchased at each period of time is very critical to cost savings. Buying too much of material will lead to high surplus and they will expire soon, buying to few of them will cause a shortage that will hinder the production. Thus it is important for the department to have a report that will collect data from all relevant sources and show them the number of weeks of production(WOS) that the current inventory is able to support.

In essence, I needed to write a set of SQL query to pull data from all sources. Using the data gathered, I had to calculate the number of weeks that the supply is able to last, and publish the visualization onto Tableau Server for the stakeholders to use them.

The result is as the following figure, for example the first material will last for 2 weeks on the 15th week of 2021. While the same material will not have enough supplies on the following week as it will have negative WOS value (demand were estimated using 13 weeks moving average).

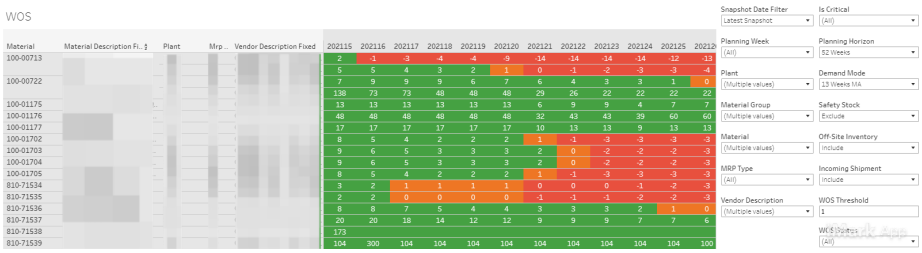


Figure 3: Weeks of Supply Report

This report needed to be refreshed periodically so that it always reflects the latest status of the supplies.

3.2.2 Report Data Sources

Supply Tables

There are two tables containing the supply information of each material — On-Site Inventory & Off-Site Inventory. Whereby On-site inventory will reflect the amount of each material-m present in the warehouse on the latest working week-i ($S_{onsite(m,i)}$). While Off-Site Inventory will contain the amount of incoming shipment that will arrive on the foreseeable weeks($S_{offsite(m,j)} : j >= i$) .

Demand Estimation

The demand comprises of three independent sources(D_1, D_2, D_3), such that each table contains the estimated demand for all materials from 3 sources.

For all demand in D_1, only demand from the present week and the following week will be taken. For all demand in D_2, the sum of demand from the past two weeks will be the estimated demand for the present week.

For all demand in D_3, ignore the demand for the present week and the following week, and only use the values from 3rd weeks from the present week.

Safety Stock

The table contains the minimum amount that should be kept in surplus for each material to be used as buffer in preparation for a sudden surge in demand.

Material Details

The table contain information that are tied to each material, such as their description, supplier et cetera.

3.2.3 Calculation

Material	Week	On-site Supply	Off-site Supply	D1	D2	D3	EOH	Surplus
001	1	100	0	20	10	0	70	70
	2	0	0	30	10	0	-40	30
	3	0	40	0	20	10	10	40
	4	0	0	0	0	30	-30	10
002	1	50	0	10	10	10	20	20
	2	0	10	5	20	20	-35	-15
	3	0	40	0	10	10	20	5
	4	0	50	10	20	0	20	25

Figure 4: Example of the final data format

The final data structure at the end will be similar to the figure. In the example, material 001 has an initial supply of 100 units on first week, and a shipment of 40 units will come in on the third week. The End On Hand(EOH) for every week is calculated by the supply minus demand for each week, but the remaining materials from the previous week can be use in the following week, hence the actual surplus the the rolling sum of the EOH value, partition by Material.

After the above table is prepared, the WOS for each material on each week is calculated using the according to the following formula:

$$WOS = \frac{EOH - I_{safety} \bullet S}{D_{x-MA}}$$

Whereby S is the Safety Stock for each material, I is the indicator variable for if Safety Stock should be considered during calculation(to be decided by the user). D_{x-MA} is the Moving Average of the demand in a window of size x (to be decided by the user).

It is clear that the value of WOS calculated will be different as and when the user chooses different MA window and the safety stock option. There are too many combinations to be considered if I were to pre-compute the different versions of WOS. Therefore the WOS is calculate dynamically in Tableau when the user chooses different settings. When an user open the Tableau report, the default will be set to 13 weeks moving average and Safety Stock as True. The user will be able to toggle the parameter in Tableau to change the computation dynamically as he sees fit.

3.2.4 Periodic Refresh

The data gathered above are stored in Database so that there is no need to repeated query every time a person is viewing the report. I wrote a Python Script to trigger the computation and write it to a another database that belongs to GPDNA. Tableau Server will then query from the table. This script was loaded onto Airflow and it will trigger the script on a weekly basis to upload the latest data.

3.3 Ariba Spend Analysis

3.3.1 Business Opportunity

Micron uses the Ariba platform for their ad-hoc purchases, where they can quickly secure the purchase of some materials that were not regularly purchased(inventorized). These purchases may be coming from different sites across the globe, and sourcing from different suppliers each time.

Hundreds of thousands of such purchases take place each quarter and billions of dollars are spent on making these purchases. It is very likely that some of materials are actually very similar in nature, and they can form a cluster. If they can identify some of these clusters and inventorize these materials, Micron will be able leverage on the sheer quantity of purchase to source for these materials more strategically. Deals can be made with suppliers for more competitive prices, stable supply channel can be established such that the shipment can be more responsive, thus lowering the risks in production and the supply of materials are much more reliable.

In the Ariba System, each material comes with a text description that the supplier uses to describe the material. Since the description is provided by the supplier, its format and content are not standardized. But it provides a fair amount of information about what the material is, hence allowing us to understand what cluster it might be forming.

Essentially, this analysis is a text classification problem, we hope to be able to group the materials into clusters through mining the information that are present in the text.

3.3.2 Ariba Analysis Stages

The Ariba analysis project is a long project, therefore I will be completing the first few stages of the project during my internship. The following sections will describe the accomplishments that I have brought to the project.

Ariba Phase-I

In Phase-I of the project, I was trying to prove the concept that the text information were indeed useful in producing clusters. Since there were no labelled data, the problem was basically an unsupervised text clustering problem.

A simple approach I used at the start was to use the processed text to generate a frequency matrix. Using the frequency matrix as input, I used cosine similarity scores and KNN algorithms to define the similarity between any two sentences. But given the fact that we have nearly 200k distinct sentences. Computing the of the similarity consumes enormous amount of space and time.

I have also tried different approach such as using the TF-IDF matrix. I increased the minimum document threshold such that a token has to appear sufficiently frequent to be part of the matrix, which reduces the size of the vocabulary significantly, and therefore the making the algorithm runs faster as well.

A problem that troubles me the most during Phase-I was the evaluation the models. Because the data was unlabeled, there are no validation or test that we can conduct to evaluate the performance of the model. With no other options, I picked a few distance-based metrics to measure the "effectiveness" of the models (Silhouette Score, Dunn Index, Calinski-Harabaz Index, Davies-Bouldin Index). Although these metrics are all distance based but each of the metrics has its own emphasis, hence the assessment of the model from the metrics often contradicts one another. Therefore, other than using these metrics, I have also consulted a few colleagues to compare the models and choose a reasonable result by looking at the random samples of the result.

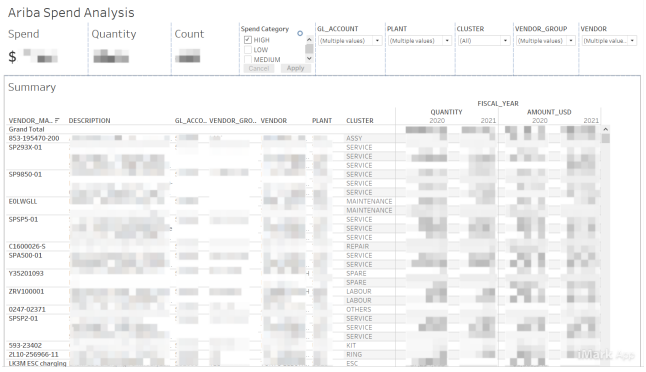


Figure 5: Analysis Report for Ariba Spend

During the first stakeholder meeting, the result were presented to the team for feedback. In the result that was presented, some of the clusters seems to be reasonably formed. My mentor commented that many of the clusters contain materials that are similar in nature, which is an indication that the model is effective to some extend, but there were also many irrelevant materials being assigned to the same cluster. A senior manager had also commented many other clusters formed did not have much business values, because there are simply too much noise in each cluster. They will not be able to use the result if there are too many irrelevant materials in a cluster.

Although the feedback for the result in phase-I was that the outcome was largely low in business value. It does, however, accomplished the intention of Phase-I, that is a proof of concept that the text description has the potential to help us cluster the materials into groups. From the result, we can gain some confidence

that if we are able to build upon the current progress and make improvements, we can definitely improve the accuracies of the model.

Ariba Phase-II

Gathering the experience from Phase-I, I reckoned that one of the main reasons that the result was poor was because Phase-I was an unsupervised problem, there was no or low basis of verification for the results that was produce. Without a basis of verification, I could not improve the models to achieve gradual improvements on the model. Therefore, I needed to find a way to evaluate the models effectively, in order to choose the best models.

Another reason the model did not predict well was because the data was too noisy. Some descriptions may contain some words that are common in many descriptions. However, it may not necessarily mean that this word is a particularly strong in predicting the clusters. Therefore I need to create a filter to remove words that are noisy, and such that the model should only take in words with strong predictive power as input.

In view of the above conclusions, some significant changes was made to Phase-I, and I call it Phase-II of the project.

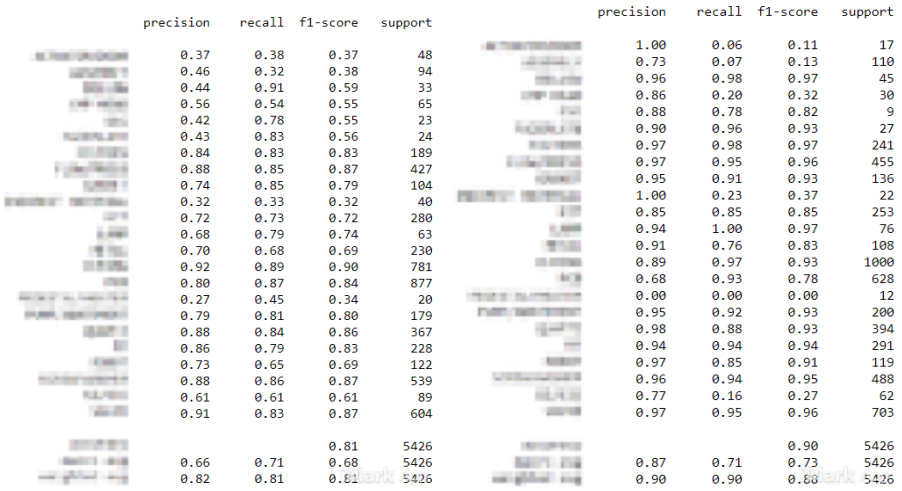
In order to solve the first problem, the best solution is to convert the problem into a supervised classification method. Which will then require a set of labeled data. Manual labeling of the materials was not an option because I do not have the domain knowledge to label them accurately. Asking others to label them for me would be too costly in term of man hour spent doing the labeling. Fortunately, procurement has another set of text data(Material Characteristics) that comes with a "L3" label(35 distinct labels). To some extend, the new text data is similar to our original data, because it is also a description text for the materials, but the text data comes from a different domain.

Under the weak assumption that the description are similar in their predictive power for L3 labels. This problem can become a supervised multiclass classification problem. I can train the model using the Material Characteristics description against target variable, L3. Then predict the L3 labels of the Ariba Materials using the model.

To solve the second problem, I need a way to be able to extract the key words associated to each cluster, and use those keywords as the model input only. (The details of the procedure taken to extract the keywords will be described in the **Learning** section.)

There were 35 distinct labels in the dataset, but because the some of the labels has much smaller entry present, the minorities in the dataset was also introduc-

ing noise to the prediction. From the business' perspective, those small clusters will not be able to provide significant savings in cost, therefore removing those labels from our target does not compromise our objective. Hence, I removed labels that do not have more than 100 entries present in the dataset, and was left with 23 labels.



	precision	recall	f1-score	support
	0.37	0.38	0.37	48
	0.46	0.32	0.38	94
	0.44	0.91	0.59	33
	0.56	0.54	0.55	65
	0.42	0.78	0.55	23
	0.43	0.83	0.56	24
	0.84	0.83	0.83	189
	0.88	0.85	0.87	427
	0.74	0.85	0.79	104
	0.32	0.33	0.32	40
	0.72	0.73	0.72	280
	0.68	0.79	0.74	63
	0.70	0.68	0.69	230
	0.92	0.89	0.90	781
	0.80	0.87	0.84	877
	0.27	0.45	0.34	20
	0.79	0.81	0.80	179
	0.88	0.84	0.86	367
	0.86	0.79	0.83	228
	0.73	0.65	0.69	122
	0.88	0.86	0.87	539
	0.61	0.61	0.61	89
	0.91	0.83	0.87	604
			0.81	5426
	0.66	0.71	0.68	5426
	0.82	0.81	0.81	5426

	precision	recall	f1-score	support
	1.00	0.06	0.11	17
	0.73	0.07	0.13	110
	0.96	0.98	0.97	45
	0.86	0.20	0.32	30
	0.88	0.78	0.82	9
	0.90	0.96	0.93	27
	0.97	0.98	0.97	241
	0.97	0.95	0.96	455
	0.95	0.91	0.93	136
	1.00	0.23	0.37	22
	0.85	0.85	0.85	253
	0.94	1.00	0.97	76
	0.91	0.76	0.83	108
	0.89	0.97	0.93	1000
	0.68	0.93	0.78	628
	0.00	0.00	0.00	12
	0.95	0.92	0.93	200
	0.98	0.88	0.93	394
	0.94	0.94	0.94	291
	0.97	0.85	0.91	119
	0.96	0.94	0.95	488
	0.77	0.16	0.27	62
	0.97	0.95	0.96	703
			0.90	5426
	0.87	0.71	0.73	5426
	0.90	0.90	0.88	5426

Figure 6: Without Keyword extraction(left) & With keyword extraction(right) Page

The figure above shows the improvement in the score after keyword extraction and the model is fed with only the keywords.

After the completion of Phase-II, a second meeting was held with the stakeholders to review the result. In the second meeting, the feedback was generally positive. The senior managers were convinced that the classification of L3 was reliable. Although there were concerns about the level of accuracy that the result. Therefore, in the presentation, I split the classification outcome into a few level of confidence (Hard, High, Medium & Low).

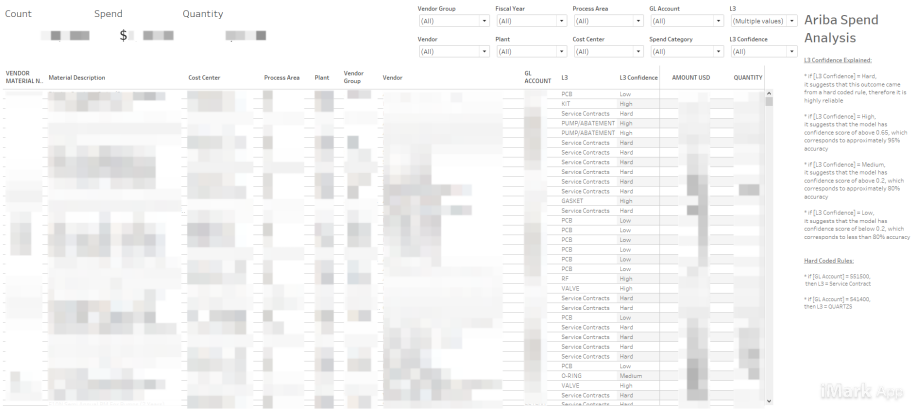


Figure 7: Result of Ariba Phase-II

Where Hard refers to a set of hard-coded rule that will be applied to find out the classification, therefore the classification is highly reliable; if the confidence is High, the outcome is correct about 95% of the time; if the confidence is Medium, the outcome is correct about 80% of the time; otherwise, the confidence will be rated low.

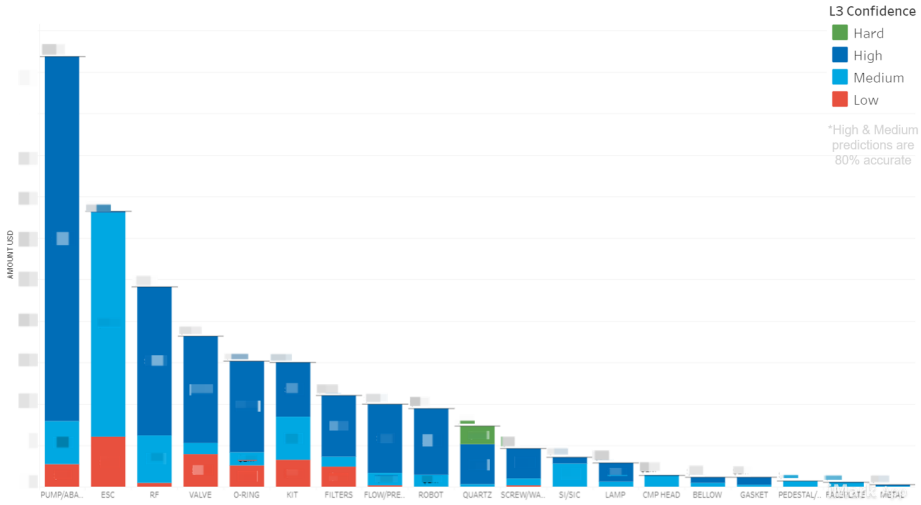


Figure 8: Confidence of the outcome

The figure above illustrates the percentage of reliable results that are present in each label. Among all the predicted classes, there are a few classes that mainly comprises Low confidence outcome, for those classes, they were excluded from results as those outcome are most likely not reliable enough for business use at this moment.

The amount of expenditure on the Ariba are in the scale of Billions, and the classification model were able to identify the L3 label for approximately 65% of the expenditure made. At this juncture, Phase-II of the Ariba project has accomplished its objective as we were able to classify the materials into their respective clusters, so that the information can be useful to other teams in Global Procurement who will be managing these materials.

Ariba Phase-III

After Phase-II of the project, the following step was make incremental improvements to the solution so that we can gradually increase the percentage of the total expenditure covered(previously at 65%). Among the 35% of expenditure that were unaccounted for, slightly less than 10% were materials with description in foreign languages, which was the reason why that they were removed from the data scope initially. The rest of the 25% that was missing was because of the removal of minority class and due to keyword extraction. During the process of keyword extraction, the algorithm look for keyword in the description and feed to the classifier. But if the description does not contain any keyword, then the description will not be classified.

In phase-III, we want to focus on including the non-English text into the solution. In order to do that, I needed to translate the foreign languages into English because the model were trained in English texts. In order to do this, I need a reliable long-term tool that can perform translation accurately on a large scale, preferably free of course.

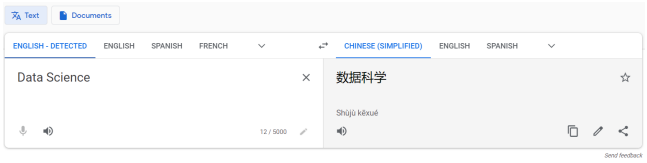


Figure 9: Google Translate Website

Of the industry leading tool would be the use of Google Translate API whereby you can submit text to Google for translation, at a charge of course. After some research, I notice that I can make use of the free Google Translate service that is on the website. Although the google website translation service can only translate a small paragraph at a time. But they have a service to upload a document for translation. So I did was to compress the text I needed to translate into a document, upload for translation, download the translated text, then parse it back into the dataframe that I need. In this way, although the process is not fully-automated, we are still able to translate our text reliably using Google's services.

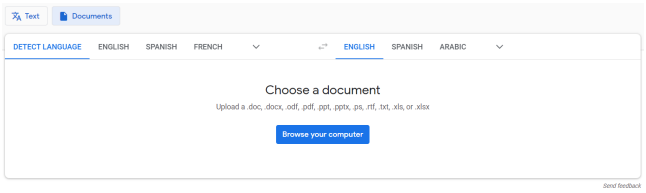


Figure 10: Google Translate file upload

With the inclusion text in other languages, we were able increase the total amount of expenditure accounted for by about a hundred million dollar. After some analysis, I realized that after the translated text was fed into the model, many of the entries were still excluded from prediction. This was because the model was unable to find any keyword in the translated text to be fed for prediction. As a result, the translated text was again ignored. Hence, we need a better more flexible model that can help to classify the materials even if they may not have any immediate keywords present.

Future Phases

At the point of writing this report, the project is ready to move on to the next phase, to explore AutoML to help solve our problem. The stakeholders are also inclined to look into unstructured data such as PDF documents used during transactions to mine more information from the expenditure made.

4 Learning

In this section, I will elaborate on details of the things I have learn from the internship and how I have applied what I learn from Data Science major to the projects. There are mainly parts in this section — Technical Knowledge & Transferable Knowledge.

4.1 Technical Knowledge

4.1.1 Docker

In the first project, I certainly learned how to use Docker.

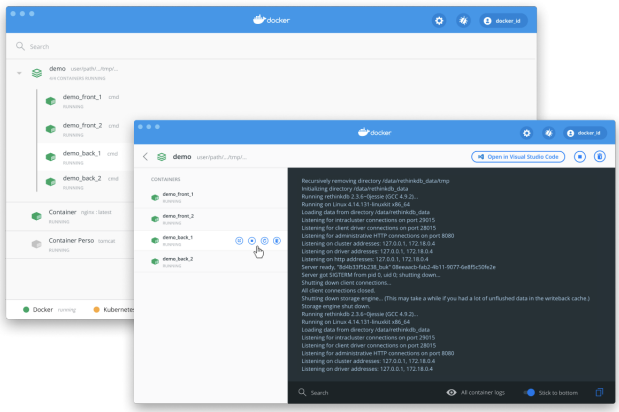


Figure 11: Docker UI

Docker is a fairly popular tool in the developers' community to host their serivce in light-weight virtual machines called Containers. However, for the Containers to host different services, the environment has to be configured using scripts called Dockerfile. A Dockerfile will specify the operating system that the Container will run with, the program that it will host, the command that it shall run when the Container when it is booted and everything else that can be configured for the Virtual Machine.

After experiementing with Docker, I set up a few Containers that run services such as Airflow, PostgreSQL, Celery(for parallel run). If it is not for the internship, I would probably not come into using Docker. After using it, I have come to understand why Docker is so popular. Because it is truly convenient and efficient in starting as many services as you in just one command. The

Dockerfile works just like a programmed code, once executed, the services will start to boot up accordingly.

The additional advantage of using Docker is the replicability of projects. Once the project is hosted on Docker, we only need to share the docker file to replicate the environment for the service to run. The best thing here is that, only the environment is replicated, while the data used to run the service can always stay isolated, if you want it to be. As such, Docker is an excellent tool to collorate on big projects, running multiple services that always need to be shared with others. The projects can be easily shared without sharing data. Which can be very critical when the data is sensitive.

4.1.2 Airflow

Airflow is another tool that I have picked up during the projects. Airflow is useful in helping to manage scheduled tasks in large scale. I have certainly not make use of Airflow to its fullest potential yet, but theres already a lot of things that it is able to do for us.

I have learned that Airflow is not limited to triggereing tasks using time. It is also capable trigger tasks using other conditions. For example, we can define a scanner that scans for the presence of a particular file. A task can be triggered once a particular file is detected. With this, we can run some tasks even if we dont know what is the time that we need to run it. Airflow will help us run it as long as the conditions are met.

4.1.3 SQL

In IT2002 — Database Technology & Management, I had the chance to learn about SQL and how to manage a database. I have already know how to use the language before attending the module, but the module has taught me about how to design an efficient database with the most approriate design. The module has also taught me some theory in constructing query using Relational Algebra, which has enhanced my understanding in SQL.

During the internship, I have the opportunity of using many of the database management systems that GPDNA have, such as HANA SAP and SQL Server. I am glad that the module has adequately prepared me for what I may need in the working environment.

However, I feel that DSA course should also prepare the students to be truly ready for "Big Data". It is very common that we have to query and calculate data in massive scale. It is very often that the query can become very complicated and the time taken could be extremely long. There werer a few situations

whereby I needed to learn to optimise my query so that I can retrieve the same set of information in a smaller amount of time.

I think it will be very helpful if DSA students are educated on the query logic in SQL and how we can optimise our query so that we can always handle data in large quantity.

4.1.4 Big Data

During the internship, I have came across many times that I had to deal with data with Billion of rows. Extraction and computation become noticeably longer than what I usually experience when I handle data in past in school.

In school or on Kaggle competitions, the data were always provided conveniently in csv format and relatively small in size. I am sure that DSA course has taught me well in handling these kind of data as most of the courses does not deal with truly BIG data. I know that we have modules such as Big Data Systems to learn about that, but I strongly recommend that DSA students should big introduced to those concepts and tools in Big Data much ealier than level 4K modules. Some breif exposure to the handling of Big Data at year 3 or even year 2 would definitely have helped me in many of the situations.

4.1.5 Data Analytics

In DSA2101 — Essential Data Analytics Tools: Data Visualisation, I learned a lot of data processing technique using R language and visualization fundamentals in GGplot. The module has benefited me greatly. What I have learned in the module is very useful when I conduct my day-to-day data pre-processing. I methodology are highly transferrable, even though I am using Python more than often.

In terms of data visualization, I have also find that GGplot is way more intuitive than Matplotlib. It allows me to focus on showing the data and leave many of the details to the Grammar of Graphics. Nevertheless, life is short, I will use Tableau for many of the visualization.

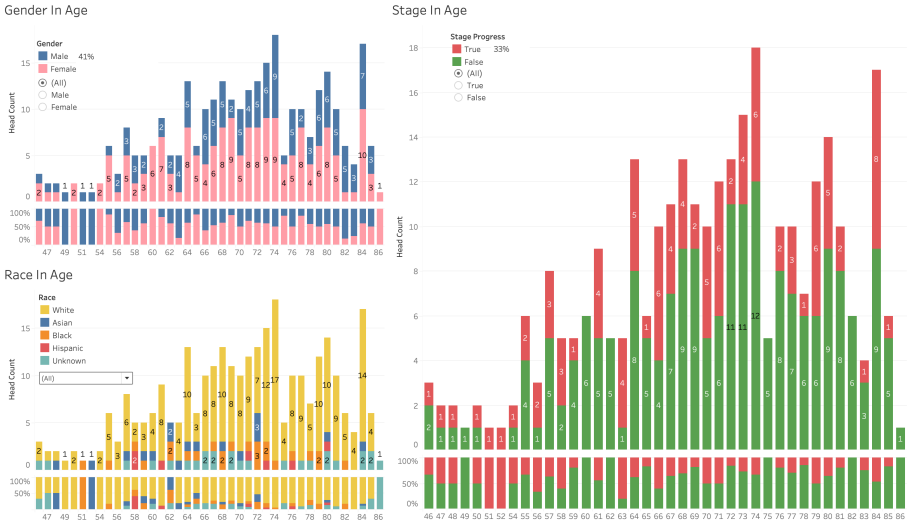


Figure 12: Tableau Visualisation

The drag and drop feature of Tableau make it really easy and fast when you need to generate a many visualization, this allowed me to visualize the data very quickly, we can then spend more time in other stages of a analysis such as the cleaning. Another amazing feature that Tableau have is the interactive visualization. Whereby the visualization can change instantly and focus on different levels of detail depending on your setting. This is very important to data scientists as they often have to study the data from different dimensions and levels.

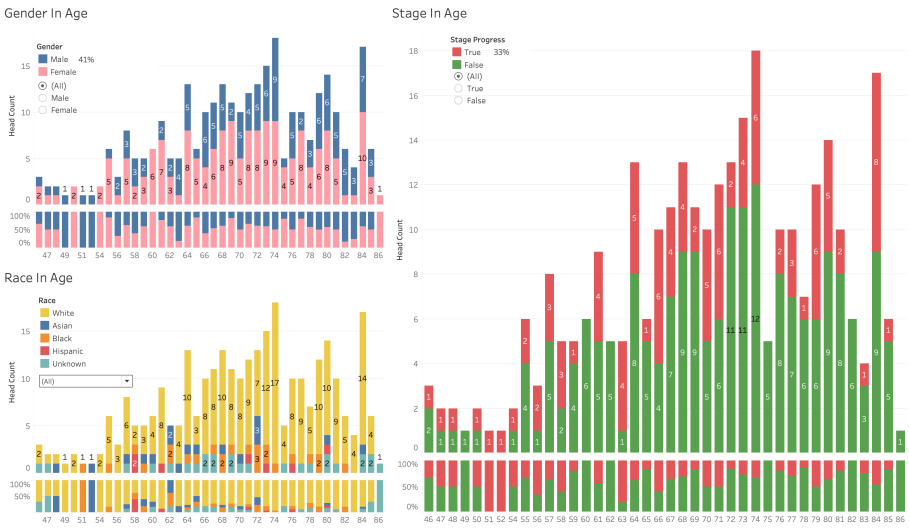


Figure 13: Interactive Tableau

I would definitely recommend DSA2101 to place more emphasize on the use of Tableau or other similar tools. Because visualization and communication of ideas is also very critical to becoming an effective data scientist. If there are tools such as Tableau that can help us communicate better with our stakeholders, then we should certainly consider investing in the tool.

4.1.6 Text Pre-Processing

In order to conduct the analysis on the text description, I have learned the different techniques in text pre-processing and cleaning. I have also implemented new techniques such as keyword extraction and Word-2-Vec.

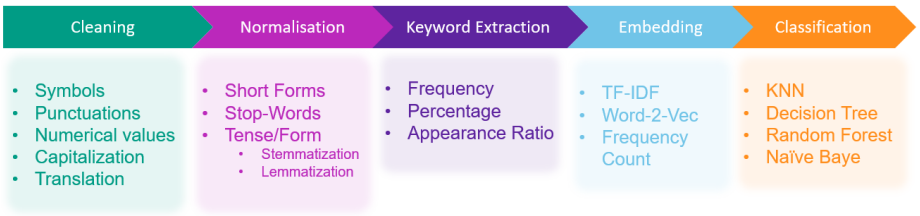


Figure 14: Text analysis roadmap

When cleaning the description texts, I first de-capitalized all the text so that the word tokens can be treated indiscriminately. I fully understand that changing the case of the characters may sometimes change the meaning of the word(apple is a fruit; Apple is the phone company). But since we do not need to process the meaning of text, therefore the above-mentioned problem is less of a problem since I will not be trying to use the algorithms to interpret the meaning of the descriptions.

After changing the case, it is followed with the removal of punctuations, symbols and numerical values. While most of the text in the data source was in English, there were about 20% of the text rows that were in foreign languages such as Chinese, Japanese etc. At the beginning, I simply remove the foreign characters to simplify the process.

I have also learned that it is also important to normalize the text just like we have to normalize numerical values in other data. Except that when you normalize a text, we will remove the stop-words, Stemmatize or Lemmatize the text.

Original	Stem	Lemma
Is	Be	Be
Are	Be	Be
Am	Be	Be
Plays	Play	Play
Playing	Play	Play
Player	Play	Player

Figure 15: Stemmatize VS Lemmatize

Stemming and Lemmatizing can often be confusing as they are rather similar. Stemming convert the word to its root regardless the form or tense. While Lemmatizing turns each word to a standardised form while making sure that each word still retain its original meaning(for example "Player" remain as "Player" instead of becoming "Play").

After normalizing the text, I needed to extract the keywords that are associated to each cluster.

In essense, I try to generate a score for each word, for each cluster(idea inspired from TF-IDF).

let w $\in W$, be a word in the set of all words
let c $\in$ C, be a class in a set of all Classes
let $S_{w,c}$ be the score of importance for word w, in class c
A word w is considered significant to class c, if $S_{w,c}$ is rank high in $\{S_{w_1,c}, S_{w_2,c}, \ldots, S_{w_n,c}\}$.

Group	Sentence
A	Today is a test for Aaron
A	Today is also a test for Aaron
A	Today shall be a test for Aaron
A	…
B	Today can be a test for Banner
B	Today could be a test for Banner
B	I need a test for banner
B	…

	Token	Group	Freq	Prop
0	Today	A	20	0.9
1	strong	A	50	0.2
2	Aaron	A	20	0.9
3	smart	A	20	0.9
4	Today	B	20	0.9
5	strong	B	1	0.1
6	Banner	B	20	0.9
7	smart	B	20	0.9
8	Today	C	20	0.8
9	strong	C	2	0.1
10	smart	C	0	0.0
11	Charlie	C	25	0.9

Figure 16: Text Statistics

In the example above, from a list of text, we can generate a set of statistics about the frequency of appearance(Freq) and the proportion of rows that a word has appeared in(Prop).

	Token	Group	Freq	Prop	OutGroupProp	PropRatio	Score
0	Today	A	20	0.9	0.85	0.85	0.69
1	strong	A	50	0.2	0.10	0.66	0.02
2	Aaron	A	20	0.9	0.00	4.50	3.64
3	smart	A	20	0.9	0.45	1.38	1.12
4	Today	B	20	0.9	0.85	0.85	0.69
5	strong	B	1	0.1	0.15	0.28	0.00
6	Banner	B	20	0.9	0.00	4.50	3.64
7	smart	B	20	0.9	0.45	1.38	1.12
8	Today	C	20	0.8	0.90	0.72	0.46
9	strong	C	2	0.1	0.15	0.28	0.00
10	smart	C	0	0.0	0.90	0.00	0.00
11	Charlie	C	25	0.9	0.00	4.50	3.64

Figure 17: Text Scores

OutGroupProp($OGP_{w,c}$) was calculated by finding the average of all **Prop** value for that word in every other class that does not belong to class c:

$$OGP_{w,c_i} = \frac{1}{|C|-1} \sum_{c_j \in C, c_j \neq c_i} Prop_{w,c_j}$$

PropRatio($PR_{w,c}$) was calculated by the finding the ratio of OGP to Prop. The ratio will be high if the word has an appearance significantly highly than its appearances in other classes. A suitable ϵ term is added to make sure that there is no zero division, ϵ can be adjusted for different cases.

$$PropRatio_{w,c} = \frac{Prop_{w,c}}{OGP_{w,c} + \epsilon}$$

There are a few typical situations: Keyword

- a word appear frequently in many rows(**Aaron** is important to group A)
- a word can be important to multiple classes

Non-Keyword

- a word appear many times, but in very few rows(**strong** appeared 50 times, but only 20% of all the entries)
- a word appear fewer times, but in many rows(**today** appeared only 20 times, but 90% of the entries)

Once these values are calculated, I can calculate a score based on the different statistics. I use the following formula for the calculation of **Score**:

$$Score_{w,c} = Prop^2_{w,c} \times PropRatio_{w,c}$$

The computation of score can always be changed to other formula depending on whether we want to place heavier weights on the frequency or the proportion of words appearing in a class.

With this technique, I can set a threshold to filter words with scores that are too low; or alternatively I can choose to keep only the top few words. I was able to reduce the size of the vocabulary to only those are significant. We can see from Figure 17 that the names are successfully identified as the keyword to each class; while the "Smart" also has a higher score to class A and B, but not to C.

Using the new vocabulary set, I can generate different embeddings such as TF-IDF, Frequency Count or using Word-2-Vec.

Text	Apple	Banana	Favorite	Fruit	Hi	Like	nice	orange
'Hi I like Apple'	0.45				0.7	0.55		
'I like Apple as well, but I also like banana'	0.32	0.51				0.8		
'Orange is so nice!'							0.71	0.71
'My favorite fruit is Apple'	0.41			0.64	0.64			

Figure 18: TF-IDF

Term Frequency-Inverse Document Frequency is a method to vectorise a document of words. For each value of word will represent the relative importance of a word to a specific document(row). However, TF-IDF is not applicable to situations whereby multiple documents(row) may belong to the same category. Therefore, I came out with the key-word extraction method above by adapting TF-IDF.

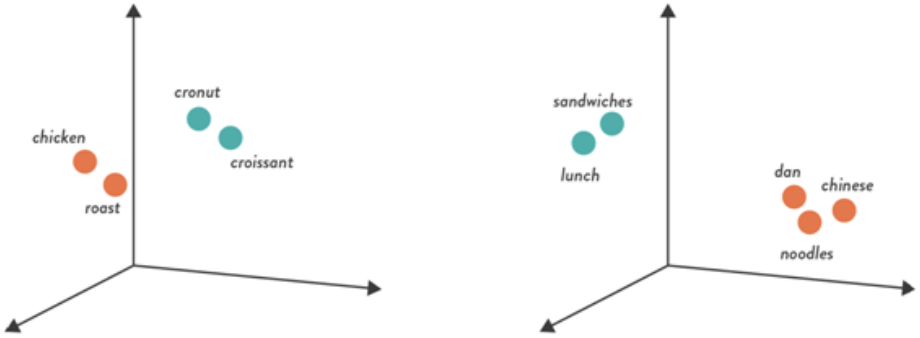


Figure 19: Word-2-Vec

Word-2-Vec is a package maintained by Google to vectorize each word into a numerical vector such that words that are related are represented with similar vectors.

Using the numerical vectors, the data can be fed to different models for training and grid-search.

4.2 Transferrable Skills

Other than the technical skills that I have picked up and applied during the internship, there are few other transferable skill. These are the skills that were not taught in school and I have leant them on the job. In this section, I will elaborate on some of the the transferable skills I found might be useful in the future.

4.2.1 Efficient Working Style

In GPDNA, the team is very special because the team comes from many parts of the world in different time-zones. It can be a struggle for some of us to work together sometimes.

This internship gave me the opportunity to adjust to this kind of working environment. This is very important as no matter where we work in the future, we might have work work with people across the globe, or the client might be from the other side of the world.

When you cannot meet with your colleagues face-to-face, it is very important to make sure that we convey our idea more clearly than ever because there will always be some messages that will be hard to convey through Zoom. It is very often that we have to coordinate our time with colleagues in other time-zones to work together. For example, when working with colleagues in US, we might have to work through our solutions thoroughly and prepare all the possible questions for the meeting. We wont be able to simply ping them and expect an immediate reply, because they will be sleeping while we are working in our noon time. So every meeting counts and we have to make sure that they meetings are as efficient as possible.

4.2.2 Business Mindset

Data Science is a profession that require a diverse set of skills. Business acumen is one of the important skills. Because it is the data scientists' job to turn data into information, but these information would be useless if they do not have any business value to it.

Having this business mindset is very important to data scientists and we must always think ahead of the value that an idea can bring to the company. As a student, I have often dwell in the technical details and excited about some of the "cool" things that we can do with Machine Learning. It is cool that we can predict some target variable accurately, but if the target variable is useless, it would be a waste of the company's resources to spend time working on it.

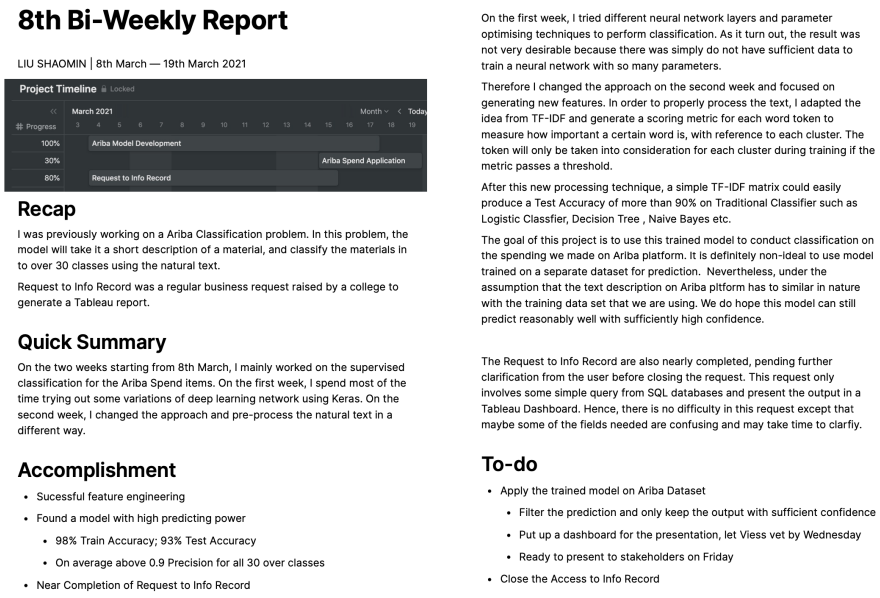
In Micron, I have often heard of the question "What is the business value to this?" The question was not just posed to me, during a meeting with the stake-

holders. I heard of the Senior Managers asking themselves for the business values of an idea that they had. I was left with a great impression of how important it is to always make sure that there is a value associated to the result that we can produce.

4.2.3 Business Reporting

During the internship, I have also learned to make report in business style. Each week I write a short report to update my supervisor on my progress during the week and I submit a report to my Academic Advisor every fortnightly for DSA4299.

The first few reports that I have made was simple and brief. My supervisor gave me some very useful tips to always make sure that a report is informative enough while making it simple.



8th Bi-Weekly Report

LIU SHAOMIN | 8th March — 19th March 2021

Project Timeline Locked

March 2021 Month Today

Progress 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

100% Ariba Model Development

30% Ariba Spend Application

80% Request to Info Record

Recap

I was previously working on a Ariba Classification problem. In this problem, the model will take it a short description of a material, and classify the materials in to over 30 classes using the natural text.

Request to Info Record was a regular business request raised by a college to generate a Tableau report.

Quick Summary

On the two weeks starting from 8th March, I mainly worked on the supervised classification for the Ariba Spend items. On the first week, I spend most of the time trying out some variations of deep learning network using Keras. On the second week, I changed the approach and pre-process the natural text in a different way.

Accomplishment

- Sucessful feature engineering
- Found a model with high predicting power
  - 98% Train Accuracy; 93% Test Accuracy
  - On average above 0.9 Precision for all 30 over classes
- Near Completion of Request to Info Record

On the first week, I tried different neural network layers and parameter optimising techniques to perform classification. As it turn out, the result was not very desirable because there was simply do not have sufficient data to train a neural network with so many parameters.

Therefore I changed the approach on the second week and focused on generating new features. In order to properly process the text, I adapted the idea from TF-IDF and generate a scoring metric for each word token to measure how important a certain word is, with reference to each cluster. The token will only be taken into consideration for each cluster during training if the metric passes a threshold.

After this new processing technique, a simple TF-IDF matrix could easily produce a Test Accuracy of more than 90% on Traditional Classifier such as Logistic Classfier, Decision Tree , Naive Bayes etc.

The goal of this project is to use this trained model to conduct classification on the spending we made on Ariba platform. It is definitely non-ideal to use model trained on a separate dataset for prediction. Nevertheless, under the assumption that the text description on Ariba pltform has to similar in nature with the training data set that we are using. We do hope this model can still predict reasonably well with sufficiently high confidence.

The Request to Info Record are also nearly completed, pending further clarification from the user before closing the request. This request only involves some simple query from SQL databases and present the output in a Tableau Dashboard. Hence, there is no difficulty in this request except that maybe some of the fields needed are confusing and may take time to clarfiy.

To-do

- Apply the trained model on Ariba Dataset
  - Filter the prediction and only keep the output with sufficient confidence
  - Put up a dashboard for the presentation, let Viess vet by Wednesday
  - Ready to present to stakeholders on Friday
- Close the Access to Info Record

Figure 20: Sample of Report

A good business report should always provide a brief overview of the progresses in different projects. Therefore a time-line will be helpful for the reader to know the progresses. A Recap will help the reader catch up with what happened in the previous report. I will add the Quick Summary and Accomplishment to describe the things I have done in the past two weeks. Finally, I will write up the full detail of the projects and what I have plan next.

It is important to make sure that the report is detailed and easy to absorb at the

same time. However, it is quite challenging as the two ideas are quite of opposite to each other. But I have learned that there are always ways to structure my report to make it more reader-friendly.

5 Reflections

In my opinion, there are still many areas of GPDNA that I have yet to discover and experience. While the team is very invested in data governance and reporting, I know that there are a lot of opportunities for me to initiate different data science projects if there are suitable applications. The team has many experienced data engineers and analysts who has been very helpful during my internship projects.

After interacting with many of the colleagues, I planned to spend a few years to gain sufficient working experience before pursuing higher education in data science, to finally work as a data scientist. I feel that this job is aligned with my career directions.

I am very happy that Micron has extended her return offer for me to work with them upon my graduation, and I have gladly accepted the offer. My skill-set is compatible with GPDNA's responsibility and there are plenty of opportunity for me to expand my skill-set. I have enjoyed doing my internship with GPDNA, and would certainly enjoy working in GPDNA in the future too.