IOT: Standup Evaluation 4
Team Noosa

1. Unsupervised Learning Summary

To increase our profit and improve the efficiency, we intended to use unsupervised learning to discover underlying patterns in our data. For example, we planned to cluster the users by the demographic data and the order profit, aiming at finding out the group of users which can bring the most profit. Also, we planned to cluster the users by the demographic data and the order items to discover the potential patterns in product preference of users. Thus, we can better recommend products based on the similarity of users in the same cluster.

However, the result of our analysis is not encouraging. We did not find any clear pattern from either K-Means or HCA. We mainly had three findings:

- The noise in the data would disturb the learning process to discover data patterns. Thus, we decided to use SSE to find the data set with the minimized error. As a result, although we planned to run K-Means on the user data set, we chose not to do it because it has a high error, which makes the result less not reliable.
- We used HCA to try to find the group of users who can bring the most profit in our business context. In our expectation, user characteristics should have some relationship with their purchasing preference. However, we did not find any interesting pattern from our result.
- We used K-means on the car-shipping data set. It indeed produced some clusters. However, the grouping result did not match to any of our expectation. Moreover, we did not find a proper explanation of the result, so we could not draw any valuable conclusion from it to improve the system efficiency.

2. Challenges

2.1 Data Collection Challenges

To mitigate the impact of the noise in the data, we need a more extensive data set than the current one. However, until now, our data set is based on 150 completed orders, which is still not enough for data analyzing, but collecting data is time-consuming. Moreover, another potential issue is the inconsistency in the data set. In the standup evaluation 2, we completed 50 orders in about 1 hour with 4 workers loading/unloading and one controller to monitor the administration application. However, in later operation activities, we only assigned 2 people to work on it. The differences in operating the system may lead to inconsistency in the data records, making it hard for us to analyze.

2.2 Limitation of the simplified business model

Another possible issue may come from the simplified business model. In the class, our data source is generated by a preset program. Thus, the pattern of the data is defined by the program. Even though we can inject some patterns into the data, it is impossible to simulate the real market. Especially when we integrate different data sources, the pattern of one feature may eliminate the pattern of another. For instance, even if there is a pattern with the product quantity in each order, it might be hidden when we calculating the total profit for each.