# Data Mining Report

# Time-series data analysis

Time-series data analysis is a popular and important statistical method. It deals with time-series data or trend analysis. It means data is in a sequence of time periods or intervals.

# Time series data prediction or forecasting

In classical statistical handling, making predictions about the future is known as an extrapolation of time series data (Mills, 2019). Forecasting entails taking models to fit on previous data and predict the future observations based on that.

# The specific method

I have used SARIMAX to predict the final result. It is used for time series prediction and stands for Seasonal Autoregressive Integrated Moving Average (Fathi et al., 2019). It is created especially for univariate time series data and supports both autoregressive as well as moving average elements. By this model, I have created an empty equation model. Fit the model with the data plus different necessary attributes. Then I predict the ultimate forecasting result for the next 29 steps or days.

# Training method

For training, I have split the actual data. I have taken the first 100 rows for training the model and put the rest 18 columns for testing the model. After splitting, I have created a dataframe and store it onto that. Then I have passed the training data into the model and forecasted the result.

# Final evaluation results

I have stored the final result in a text file of the project folder. The first column of the file represents the key product ids. From the second column onward, it is showing the forecasted values for the next 29 days. Each column represents each day’s forecasting value up to the 29th day. As I have also shown the predicted output using matplotlib at the end of the notebook.

# The relationship between the minimal amount of training data and the prediction accuracy

I have less data sample for this project, my model learned less. Based on the limited learning, it predicted the final result. If we provide more data into the model, then it will be capable more to predict the value with more accuracy. We will never get the best accuracy if we use a linear model to categorize data that is distinguishable in a nonlinear way. So, finally, we can say that the prediction accuracy is always based on the sample of the data. The minimal amount of training data predicts as per the limited learning.

# References

Fathi, M. M., Awadallah, A. G., Abdelbaki, A. M., & Haggag, M. (2019). A new Budyko framework extension using time series SARIMAX model. Journal of Hydrology, 570, 827-838.

Mills, T. C. (2019). Applied time series analysis: A practical guide to modeling and forecasting. Academic press.