Module 2 Assignment XN Project Digging Deeper

Esha Mulki

Mohammad Movahedi

Ajoy Kumar Nandakumar

Taiye Murtala

College of Professional Studies, Northeastern University

ALY6080: Integrated Experiential Learning

Dr. Matthew Goodwin

October 9, 2022

Abstract

This paper is about the exploratory data analysis for the Danfoss dataset. The Danfoss dataset is first prepared for analysis by cleaning the dataset. The dataset mostly consists of numerical values with 174 rows and 24 columns. The columns were renamed to have clean names and missing values were removed. The descriptive statistics of the dataset was checked for the mean, median, maximum and minimum values. Outliers were checked and it was found that columns like ‘emea_crude_oil_prices’, ‘emea_pmi’, ‘emea_vdma_machine_building’ had many outliers. From the correlation plot it was observed that some columns were closely related to the target variable which are further analyzed. Graph was plotted of the target variable with respect to the years to check the trend pattern of the target variable from the year 2008 to the year 2022. This graph shows that there was spikes in the target variable for the year 2008 and 2021. Graph was also plotted of target variable with respect to the months and it was found that there was not much variation in target variable with respect to the months.

How are you addressing data preparation?

The dataset is prepared using the R language. Libraries such as "dplyr", "tidyverse", "ggplot2", "corrplot" and "janitor" are used to gain an understanding of the data on hand and clean errors and accuracy along the way. Analyzing the dataset, it is observed that most of the data are numerical values. The dataset contains monthly data from the year 2008 to 2022. Some of the columns 'Employment Rate', 'Production of total manufactured intermediate goods Index',' Production of total manufactured investment goods Index' and 'Residential Property Sales of Newly Built Dwellings' have missing values, sourcing these missing values needs to be discussed with the client. The descriptive statistics of the data was performed where the mean, median, maximum and minimum was checked. Outliers were found in some of the columns.

What tasks did you complete?

On plotting the correlation plot it is observed that more that five variables have high correlation with the target variable which means any changes on the variables also impacts the target variable. Graphs were plotted for the top 5 highly correlated variables with respect to the years to check their trends from the year 2008 to the year 2022. In the graph it is observed that target variables spike in the year 2010 and 2022. Also, graph was plotted of target variable against months, and it is observed that the target variable does not show much variation during different months of the year.

What tasks are left to be done?

Discussion with the business is critical for understanding information about the different columns, target variable and the missing data. Further analysis needs to be performed on the highly correlated variables. Also, the dataset is not divided into training and testing sets since the optimal machine learning (ML) approach is not finalized. If Bayesian Model Averaging (BMA) model is used, then the dataset need not be divided.

What is your plan to complete these tasks?

Online meetings will be conducted with the sponsor to get the information related to the missing data. Research more on the highly correlated variable and investigate relationship between them. In-person meeting will be conducted within the team to finalize the ML approach which need to be taken.

What are the preferred methods of communicating the results from your initial EDA?

Reports containing visualizations like graphs along with description about them will be preferred methods of communicating the results of the EDA. Also, PowerPoint presentations can be conducted for presenting the analysis insights to the clients.

How do you plan to communicate results of tasks yet to be complete?

Work with the team to determine the key objectives of the tasks, determining what data will be important for the further analyses and focus on the critical points. The process of the project can be monitored at specific time intervals to see that things are working as per plan. Meeting can be scheduled to discuss any open issues or tasks that needs any inputs from the clients.

References

Danfoss. (n.d.). *The journey to Engineering Tomorrow*. Danfoss. Retrieved October 08, 2022, from https://www.danfoss.com/en/about-danfoss/company/history/ (Links to an external site.)

Bahler, K. (2020). *Inside Amazon's Very Weird (But Very Efficient) Staff Meetings*. [online] Money. Available at: https://money.com/amazon-meetings-no-powerpoint/ [Accessed 9 Oct. 2022].

Bhanot, P. (2021). *Six Essential Data Preparation Steps for Analytics*. [online] Actian. Available at: https://www.actian.com/blog/data-integration/the-six-steps-essential-for-data-preparation-and-analysis/ [Accessed 9 Oct. 2022].

Talend - A Leader in Data Integration & Data Integrity. (2022). *What is Data Preparation? Processes and Example*. [online] Available at: https://www.talend.com/resources/what-is-data-preparation/ [Accessed 9 Oct. 2022].