Assignment 3: Assessing Data Management

By

Mehnaz Afrose

In the Data Mining and Management part, I have completed one course CIDM 6351 - Business Data Extraction, Transformation, and Load in Fall'23 semester and currently studying another course CIDM 6355 - Data Mining Methods in this Spring'24 semester.

In this reflective assessment, I will thoroughly explore my competencies, strengths, weaknesses, and potential areas for improvement, focusing specifically on the domain of Data Management. Through a detailed examination of completed coursework, samples of submitted work, and relevant sources I've consulted, my aim is to offer a comprehensive overview of my abilities and preparedness to employ this knowledge in future endeavors. This assessment holds relevance in terms of how I can contribute effectively to the forthcoming capstone project.

What do I know: competencies, skills, and knowledge:

From Business Data Extraction, Transformation, and Load:

The biggest achievement from this curriculum is I got the opportunity to achieve the Data Engineer certificate by completing required eighteen courses from datacamp.com. After this I was able to complete a data engineer project instructed by the professor. In this project I demonstrated ten data cleaning methods of the dataset of ufc fighter's statistics which I got from https://www.kaggle.com/datasets. After cleaning the dataset, I analyzed several parts, for example is there any relation between the win percentage and height of the players or between the stances. The reference of my work: GitHub_Data_Engineer_project. I also learned how to perform importing data from various sources and then clean and transform the dataset to make prepared for analysis. Also, how to analysis the data. The demonstration of my work: Homework2 and Homework3

In a nutshell, I gained the skills about data engineering by cleaning the dataset for analysis and then analyzing the data to gain insight from it.

From Data Mining Methods course:

From this curriculum I gained knowledge about data analysis process and data mining concepts, algorithms, theories, and processes. I learned to use data mining tools RapidMiner and R to implement data mining algorithms to perform various data mining

tasks on large data sets. Through each week's homework, quizzes, and lab practices, I skilled at data mining of structured data and text format data. The demonstration of my work can be found in my GitHub repository.

Areas of Improvement:

From Business Data Extraction, Transformation, and Load:

In this course, there was a project about web scraping. I needed to web scrape the four tables of data at https://guides.lib.uh.edu/OER/ATIP/awardees . Then I needed to store the data from all four tables into one pandas DataFrame with columns "Year", "Name", "College", and "Course". Sort the data by year.

In accomplishing this project, I encountered difficulties as I have very limited knowledge about web scraping. Additionally, I am weak in the Python programming language.

From Data Mining Methods course:

Most of the data mining we have accomplished so far has been for structured data, which were in a tabular form in a CSV file. There was only one text mining exercise. However, I feel the need to learn how to mine semi-structured and unstructured data.

Future: What do you wish you knew and/or don't realize you are missing:

From Business Data Extraction, Transformation, and Load:

I wish to become more proficient in the Python programming language and web scraping. That's why I am planning to review all thirteen chapters from the codewithmosh.com website. I have already become a paid member of this page.

Additionally, I plan to revisit the web scraping tutorials from the class lectures and watch video tutorials from online media platforms.

From the Data Mining Methods course:

I wish I could learn how to mine unstructured data along with semi-structured data. Currently, I have become a paid member of the IBM Coursera webpage monthly. I am planning to look for courses there about this subject.

Contribution Towards the Capstone:

I have developed a project outline for my capstone project. For this project, I plan to develop software for doctors to predict the risks of diabetes for their patients. This application aims to predict the likelihood of a particular patient developing diabetes.

Why Data Management is Integratable:

For data preparation and cleaning, in developing machine learning models or algorithms, visualizing the results, and for future analysis, the knowledge and skills from data management are integratable with this project.

How Data Management is Integratable:

The dataset I am using for model development and comparison contains 100000 rows records. It's divided into 9 attributes with one target variable called Diabetes.

For data preparation and cleaning, I used Jupyter notebook. Also, for developing the classification models, predicting the results, data management was used.