Group Project in Data Science - CSC8633
Word Count - 1351
Rohit Kumar- 220524737
Group 12

# Reflective Report

**Description**

For our module Group Project in Data Science, we were provided with the opportunity to work as a group to deliver a data science project. This module allowed us to learn various new skills, such as working in a group, and various soft skills while at the same time enhancing our existing technical skills. The objective of the project is to improve the recruitment efficiency and hiring success of HC-One, which is the largest care home operator in the UK. HC-One receives approximately 2,500-3,000 job applications per week for care workers and nurses, mainly generated through third-party websites. However, the recruitment team needs to assess each application to determine if the candidate is likely to be a successful hire, which is a slow process causing the organization to lose potential high-quality candidates. Therefore, our approach aims to analyze how the different variables provided in the datasets have an impact on the success of a candidate and based on that develop analytical and forecasting solutions to indicate the likely success of job applicants. The business understanding and problem statement align with the CRISP-DM methodology's first stage, which is understanding the business and data objectives and the context in which the project will take place.

The initial step was to gel with the group and then perform brainstorming sessions for understanding the business objectives, requirements, and Methodology to be opted and discuss the overview of the whole project cycle so that, every team member is clear about the road we are going to take to complete and submit the deliverables.

During the brainstorming sessions, several decisions were made keeping in mind the end deliverable. Few key insights and decisions made during these brainstorming sessions:

- We decided to follow a mixture of SCRUM(Agile) and CRISP-DM-inspired methodology.
- We created a project plan by following the sprint methodology to fulfill the delivery of a minimum viable product.

- We made this plan keeping in mind the flexibility to scale the product in future work.
- We also created Functional Team PODs or team leads so that we can separate into smaller units and parallelly perform different tasks while communicating with each other so as to save time and then in the end come together and create a cohesive deliverable product.
- The team pods were made on the basis of our project plan to divide the deliverable product into four parts-  Data Visualisation, Data Preparation, Modelling, and Data Insights and Analytics tasks as part of a first sprint delivery
- I was in charge of the Data Insights and Analytics team.

The next step was to start working on the project. The Data preparation pod started working on cleaning and processing the data for the other pods to use. The Visualisation team created mock-up data to perform some visualization.

I collaborated with the modeling team to help them understand the raw data and the cleaned data obtained from the Data preparation unit. When this was done the modeling team asked me to perform analytics on the cleaned CSV datafiles to gain some valuable insights and help the modeling team perform feature engineering based on my analytics and insights. The exploratory data analysis performed by me helped in gaining some visualization, out of which some were unique while few were redundant because the Visualization team was also able to gain similar and more appealing visualization they prepared using Power BI in contrast to the ones made by me using Python Programming language.

So a part of the EDA performed was a recursive reference to the analysis performed by the Visualization pod, strengthening their claims and insights with additional validation.

I was also one of the few members of the group who were selected to create the group report and presentation. Regular documentation was done in intervals so as to keep track of the group's progress so that the integration of different tasks into a single pipeline at the end of the project would be smooth and a cohesive product could be generated.

## Feelings

While previously working on a group project for a prior module, I was quite excited and prepared for the group project for this module. I was able to develop several group skills before and was confident to sail through this group project with ease. But working with a whole new group and a whole new project was overwhelming for me. This group project helped me understand the very important fact that every project is different from one another. No two projects would have the same underlying objective or requirements, no two projects would have the same group dynamics.

## Working with the Team

I learned that it was very important to communicate with different team members to smoothly work in harmony. I learned that for a team to reach its objective and fulfill the deliverables, the key factor is to bond with the team members. I achieved this by asking the team members out for a surfing session. This enabled members a bit hesitant initially to gel with the team. Following this activity, several members took the initiative to have various team engagement activities across this sprint so that the team morale is always high.

## Evaluation

Upon reflection, I think the team performed very well in fulfilling the deliverables. Initially the team, quickly got together to address the problem and find out ways to tackle the problem and create solutions. It was a bit later when there were a few conflicts of ideas, but the team worked together to overcome these obstacles. One of the specialties of our project which the whole team agrees upon is the cohesive end product. We know that reproducibility is important when delivering data science projects. As software components, we should also follow good practices in terms of modularity, testability, coupling, and cohesion. Separating this project into independently deployable units yet being cohesive allows us to be ready to follow industry practices and enables us to scale our product for future requirements.

## Analysis

In retrospect, I realized that I made some assumptions about the data that may have influenced my analysis. For example, while finding out the highest correlated variables for our dataset, I took into consideration the feature having an absolute correlation value of more than 0.9. Meaning, since I took the absolute value, it has given me both positive and negative correlated features, which I assumed would be beneficial to have. Also, there were a few assumptions made while removing duplicate candidates. For example, I assumed that the duplicate observations were created because the stakeholder did not have a definitive system to update a candidate's application by overriding the previously saved instead it created multiple entries of the same application whenever a candidate made any changes in their saved application. This may not have been the case which would have resulted in some biased analysis.

## Conclusion and Action Plan

Despite these limitations and challenges faced while working, I was able to identify some key insights that could be used to improve the product line and increase stakeholder satisfaction by helping them in hiring more qualified and suitable candidates, that can provide higher quality care to its residents, leading to better health outcomes. Following the project, I reflected with a few teammates to get feedback and understand my strength and weakness.

As mentioned earlier, the project plan was created keeping in mind that we have the flexibility to improve upon the product in future sprints. Therefore, Having modularity and deployability in mind we believe that by incorporating MLFlow into the project to implement deployability in the model, as well as better organization of hypermeter tuning due to its experiment tracking capabilities we would be able to produce better results. Also on an individual level iI believe that having multiple pods working parallelly, it might be better to have a flow so that there are no redundancies between visualization and exploratory data analysis. Additionally, I would like to be more mindful of the assumptions I make when analyzing data to avoid bias and ensure that my conclusions are valid.