INTERNSHIP AT THE CAMPUS ANALYTICS

"The only way to do great work is to love what you do. If you haven't found it yet, keep looking. Don't settle. As with all matters of the heart, you'll know when you find it."
- Steve Jobs

It's been a while that I decided to pursue the Science and Technology Internship subject (SCIE90017) at the University of Melbourne and I am very happy with my decision to work at Campus Analytics as a Data Science Intern. I realized that this is where I want to work. If I stick right to my plan, I can accomplish all my goals.

The experiences of working as a Data Science Intern at the Campus Analytics was fascinating and it gave me enough exposure to the real-world side of this field. This experience motivated me to stick by this area and do wonders in the future. This essay is a reflection of my internship experience and is broadly divided into three sections which are: Knowledge, Technical Skills and Soft Skills.

In terms of the knowledge, I learned many different things and I was brushing up with my basic information about data science too. Campus Analytics analyses the space and its usage across the entire University of Melbourne and provides the authorities with a proper solution of which area is being used nicely and which is not. My role in the internship was to compare three different libraries across the Parkville campus of the University of Melbourne. The comparison of the ERC, Baillieu and the FBE libraries was based on the space available and the number of students visiting the place daily. I was interested in producing the output which could show which library is more crowded, or less crowded or which library do students prefer more, which library needs some changes so that students visit more, etc.

Initially when I joined the company, the director and the other authorities made me aware of the office work culture. The next step was the allocation of my project. I was explained the way I need to go ahead and the different ways to get the output and convey to the end-users. The first step was to get the data and understand it. I received the data from the Data engineer at the organization. The main issue I faced here is to figure out a proper way to handle the data as the data was very huge (around 11GB) and it was very difficult to load in an IDE directly to process it. I had to get the data in two different formats: CSV and the parquet. The parquet file is in a similar format as the CSV file but with very little capacity in terms of the disk space used. Due to some system issues and big data problems, I had to work with the parquet files as the CSV files were very big and they were not feasible to use at that moment. Further, I started working on the cleaning and processing of data. I deleted the unnecessary columns, checked all the values and made them accurate as per my needs and the goal I had to achieve.

Once I got all the data in the proper format, I started exploring the data to develop some trends from it. The data consisted of the number of students visiting the library in each hour of the day, the number of students on each floor, some personal information, etc. I worked on the data and developed some trends and plotted them accordingly. Studying and understanding the data this way helped me to get a clear knowledge of the data in every aspect which made me capable of answering any question related to it. Then after, I extracted the data as per my needs and

performed mathematical operations such as average, summation, etc. on the data to get the output. In terms of the learning outcome from this part is that I learned how to go ahead when somebody gives you any data to process and you return to them with some valuable and business-related outputs. The standard process is following a series of 5 steps. They are Data Collection, Data Preprocessing, main task to be performed (Algorithms, mathematics), Data Visualization, Present to the responsible people, etc. I learned the way to start with a project in real-life scenarios and how to tackle different problems we face during the entire project and completing the project successfully at the end. Also, I am aware of all the information about different types of data handing & processing, thereby generating some business outcomes out of it. Another key learning outcomes are information on data science and why it is important in the real world and how can data science solve problems of the modern world.

As mentioned, my internship was based on the manipulation and the processing of big data. So, eventually, I had to grab a lot of technical skills and apply my previous knowledge to it. I started working on python as it has a lot of libraries using which I can do the data manipulations and get the desired results. Pandas is one of the widely used libraries in Data Science. It has inbuilt objects such as the Dataframes and series which are very useful for data analysis and processing. Pandas is majorly used for working with tabular data and in varied formats such as CSV, SQL, parquet, etc. Another important and frequently used library in python is NumPy for data science. I used this library to work with the numerical part of my data. NumPy is a linear algebra library in the python programming language. Using this library, I performed various mathematical operations on my data with the help of arrays. As NumPy has a lot of different machine learning inclusions, it's also known as the Machine learning library in python. Tensor flow for machine learning and other libraries also use NumPy internally. Handling this huge quantum of data was a major issue for me. But eventually with the help of the seniors I figured out two different ways to handle that, namely Dask and PySpark which help us to manage very huge data files. Dask works on the principle of parallel processing. It distributes the workload to different processors and then completes the operations performed. This way it reduces the time taken for the execution of the task and in the end, it merges the output from all the processors which is the correct and accurate output. PySpark is an API that is integrated into python to support Apache Spark with all its operations. Apache Spark is a framework that is used to perform big data analysis. It processes and performs operations on the data with lightning speed thus making the operations very efficient. Using the pivot table functionality in python and Microsoft excel, my work was much simpler as it could summarize the entire Dataframes, count the values, reorganize it, group, average, etc. In my project, it helped me to count the unique number of students on each floor in each library. With the help of the pivot table, we can view the data from many different perspectives and can visualize it accordingly. Pivot tables are very useful in Microsoft excel as well but to load very big data in excel (preferably up to 3-4 GB) and to apply some functions on it, there is a need for a plugin to be installed which is known as the "Power Pivot". Google Data studio and tableau are the two different software's which helped me to visualize my results and convey them to the authorities in a very proper manner. Proper visualizations and making the experts understand the results are factors that made my success in my internship.

As a data science intern, I had to learn many non-technical skills as well. During the period of this internship, I developed my critical thinking ability which helped me understand the problem statement, make hypotheses, etc. I also brushed up my problem-solving skills, intellectual skills, and developed business sense along with technical knowledge. A data scientist needs to provide a solution in such a way that it helps in increasing the business value for the project he/she is working for. One of the major skills I learned is "Effective Communication". As per the research, most of the data scientists lack communication skills. In my internship, I was told to focus more on presenting my work as the managers, end users are not interested in the technical terms. They are more interested in the output, the storyline of the entire task. Effective communication is the way in which you convey our project's results in a proper and interesting manner. This includes speaking skills, writing and presentation skills. As per the recent article on the internet, people with excellent communication skills are hired and absorbed faster by the industries in the market.

There are many challenges you must face in the professional life as compared to academic life. I faced those such challenges during my internship at Campus Analytics. The first one is the change in the social environment which seems very weird as there are very fewer colleagues with similar interests as you and it is very difficult to be on the same frequency as them. In terms of hangout, sports, and other fun activities, I had to restrict myself as there is no time available. Responsibility is the buzz word in the life of every professional. I had to complete the project on time with accurate output as it was expected. In industry, there are no other alternatives to the timely and successful completion of work, unlike university in which there is some casual attitude students follow and take things lightly. The most important change I faced in professional life is the need for proper attire and good communication skills. By these two things, people in the working environment are judged and people behave with them accordingly. If I compare the technical and non-technical side of the university and the industry, skills that I learned in the University are very useful in the industry as well. Skills learned in the University helped me develop roots for a better future in my professional life.

Based on my experience in the industry after working at Campus Analytics, there are certain things I would like to suggest to the University. Firstly, universities should upgrade their curriculum and add certain subjects that are restricted only to the basics of that domain. The reason behind this is that people face problem in the industry which occur because of the basics being not clear and the concept behind a particular thing which is not properly understood. Another thing which a University can do is make the Internship subject compulsory rather than keeping it as an elective. Unless there are some students who wish to pursue research subjects for the Ph.D. otherwise Internship can be made compulsory for students who want to get involved in the industry after the course. Overall, the Curriculum of the University of Melbourne has been very helpful to me in my internship and I have strengthened all the skills which the University taught me during this internship.