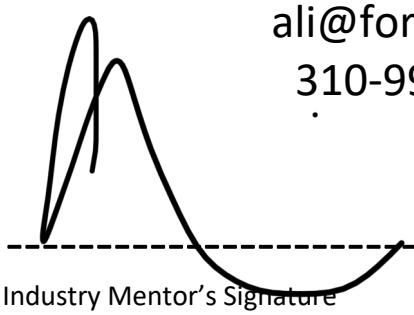
CSE 584
Summer 2020
40948

CPT Internship Report

FORKAIA

NEHA TANDON
ASU ID: 1216798336
MCS

ALI SINA
CEO
ali@forkaia.com
310-994-1511



8/14/2020

Industry Mentor's Signature Date

Faculty Chair's/Faculty Mentor's Signature Date

Neha ----Neha Tandon---- --08/12/20--

Student's Signature Date

Student: By signing you certify that you have not plagiarized any of this document.

Program Chair's Signature Date

1. **Educational objective of the internship:**

   During my internship, I analyzed datasets, interpreted the data and identified trends, patterns and useful strategies. This required me to apply the Data Mining techniques I have learned in CSE 572. I also modeled and created databases in SQL for easier search and query and supported powerful analysis through visualizations, for which I applied my Database Management Systems and Data Visualization knowledge that I have gained through my courses CSE 511 and CSE 578 during my Master's at Arizona State University.
   Working with my colleagues on multiple projects at the same time taught me time management and enhanced my organizational skills.

   Specific job responsibilities:

   - Provide the Business Team with more information literacy regarding their Startup
   - Collect, analyze, and interpret data relevant to your chosen startups into useful strategies and identify patterns and trends in data sets.
   - Formalize, model and create databases for easier search and query and define new data collection and analysis processes.
   - Conduct Research and pull data on : User Group: Summarize who your app/startup is targeting as users | Implementation Details: Outline what your app/startup would need in terms of data structures, databases, outside APIs or resources, etc. | Existing Market: Outline what types of similar products exist, what are their shortcomings, how can we improve them, what challenges do they face, etc. Analyze, Model, and Interpret data into useful strategies.

2. **Company information:**

   FORKAIA is an idea lab and innovation accelerator that's birthing companies positioned to fundamentally change existing markets. They find, build and invest in AI start-ups & early stage technology companies that are disrupting the present, in service of a better future. The FORKAIA open-sourced ecosystem consists of the most talented engineers & scientists worldwide collaborating and working on some of the most interesting and groundbreaking projects on the planet using artificial intelligence, machine learning, data science & neural networks [1].
   FORKAIA's one of a kind plug and play startup platform is where the world's greatest resources in a soon-to-be AI-driven world (young, talented human brains) are trained, drafted, managed, compensated & placed (traded) where they are maximizing the application of their skills for the betterment of humanity as a whole [1].

   FORKAIA has a collaborative environment where challenges and ideas are discussed openly. They cultivate performance & leadership through empowering team members, diversity of thought and a commitment to an inclusive culture that values hard work, focus and discipline. Everything from quick design questions, philosophical discussions, to fun ideas for our technology – nothing is off limits. We are an engineering-driven organization; everyone here makes product decisions. They are now working on over a 150+ startups and have a 5000+ Persons Development & Data Science Agency, an Innovation lab, Startup Academy, Creative lab, Accelerator, a Talent Agency, a Farm system that nurtures, trains, and manages a new type of talent – the Business Athlete. The culture at the firm is built around collaboration, experimentation and disruption. Artificial Intelligence will transform the workplace by automating a sweeping array of human activity. FORKAIA Academy is reshaping education to meet the nature, speed, and dynamics of these changes and rethinking the role and purpose of human workers in an age

of informational abundance, process automation, and machine autonomy. The Free Synergy Ecosystem gives our members access to move freely within roles, teams and startups and choose to work on whatever they want. You can open any of FORKAIA's doors and try out different companies, sectors and roles so you can get a good feel for what you like and don't like as you prepare to launch your career. We are making the next Facebook and Dropbox in our lab right now using the Collective intelligence of hyper-diverse teams. Our Team members have the utmost creative freedom to work on their passions & be able to make a significant impact early on. We want our team to think unconventionally, be creative and find elegant solutions to problems that are yet
unsolved [1].

As a part of The FORKAIA Startup Combine program, you gain experience working on solving real-time problems in your field of study with a diverse team of programmers. It's an opportunity to learn from each other, grow professionally, get valuable feedback, and enhance your professional reputations because of what you're able to accomplish together. Along with real experience, working for the FORKAIA Agency results in actual work product that you can point to and say, "I did that." Because their Agency clients are usually project-based, it is a very clear demonstration of your skill - Agency projects put project-based learning into action. This gives great exposure, showing them AND their clients our capabilities. FORKAIA's vast network and ecosystem introduces us to professionals in our field who might be valuable contacts later. This is a fantastic opportunity to brand oneself, build a reputation and more importantly build a portfolio [2].

As a Data Science & Analytics Intern, I worked as a part of the Data Team. This team is responsible for gathering data, analyzing and drawing insights, and offering business solutions. They focus on creating procedures that will help process data efficiently once the startups launch and once numbers begin flooding in. In the meantime (i.e. pre-launch), data analysts set up strategies to leverage data in the future. The Data Team has many talented members with varied backgrounds and experience, who work together on the assigned projects to discuss and implement strategies for: cleaning and consolidating raw data that is received, creating and managing databases, performing analytics and creating visualizations for reports and presentations.

3. **Background about the problem:**

I have worked on a number of tasks as a Data Science & Analytics Intern. Some of these were team tasks and some others were individual tasks. The main focus of the problem encompassing these tasks was- raw data cleaning and management, analyzing the data and consolidating the insights in the form of visualizations. The context of the problem was mainly handling the company or client data and performing and reporting analysis. This problem needed to be solved in order to give the company/client more insightful information about their employees/users, provide easier access to the data and assembling a portfolio that showcases to the clients what the company has to offer. Both the company and the clients benefitted from solving these problems as it gives the company an integrated brochure that clearly illustrates the company's talents for sharing with the clients and also helps the clients understand the characteristics of the company and make an informed decision based on their aspirations and requirements.

Using the data as leverage for future business decisions was the idea of this project, whether the data was company's employee data or the client's data. This was also extended to help provide insightful

solutions to help our clients make better decisions and use these insights for improving their products. The main target audience of a start-up organization is its users. Understanding their needs and taking their feedback into account not only helps provide better services, but also builds trust and understanding for the future. The company uses its data as stem-cells.
The team-based tasks involved brainstorming solutions with team members to determine the best approach for constructing an automated solution that can be deployed by the company to be effectively used on new data and give continuous insights that help the company make better business decisions and discover the strengths hidden within the raw data. We had to create a showcase for FORKAIA's Talent in a compelling way. Using the given data, we had to paint a visual picture of FORKAIA's Talent. We found the most noteworthy, impressive accomplishments, awards, recognition or honors of our talent to extract and highlight.
The individual tasks were solely handled by me, with guidance from my supervisor and the project manager. These tasks were based on Data Visualization. However, the data had to be first pre-processed and organized to make sense of it and use it to create these visualizations. Multiple visualizations were created by me to showcase the insights extracted from the data; these results were appreciated by my superiors.

4. **Focus of the effort:**

My approach towards the problem was task dependent. The one thing common to each of these tasks was that, the data given by my supervisor was raw data which had to be cleaned to make sense of it, incomplete values had to be corrected and missing values had to be dealt with. This constituted the first step towards the solution. The second step was organizing the data to have easy access to information. Designing the databases with the appropriate structure and storing the data in databases was a part of this step. The next step was performing analysis, to give the gist of the data. Lastly, the insights were represented in the form of visualizations to offer a more appealing and comprehensible solution. Working with my peers, I was able to actively bring ideas to the table that were worth discussing. Brainstorming as a team helped us incorporate important perspectives to come up with a more well-rounded approach for our tasks. Maintaining a timeline for our tasks and having weekly progress meetings allowed us to have a clear-cut plan to deliver the best quality of our work.

I had to learn some skills to tackle this problem, like: SQL, NLP, text extraction techniques, pattern matching in Python, Python for Data Science and Tableau for visualizations.
The course CSE 572 Data Mining helped me in dealing with the raw data. As a part of this course, I learned how to perform data pre-processing, how to deal with missing values and how to extract useful patterns from the data, practically through projects from this course. From the course CSE 511 Data Processing at Scale, I learned how to design the structure of databases, effective database management, how to handle Big Data and how to use MySQL Server for database management. The course CSE 578 Data Visualization taught me creating data visualizations and storytelling skills that played a very important role in creating meaningful visualizations for these tasks. As an analyst, understanding the data is most important to be able to showcase the data. Hence, I ensured regular discussions with the Project Manager, the Supervisor and my colleagues to get more insights about what the data represents and shared my approaches with them to incorporate their feedbacks for a satisfactory end result.

5. **Results:**

The results of my work were presented in the form of databases, excel files with key insights and a collection of visualizations, which were created from the results of the analysis of data, to provide a

better visual representation. Some of the visualizations that resulted from the analysis of the raw company data are shown below (Fig 1 and Fig 2).
Extracting these insights required learning text extraction techniques in Python like NER, regular expressions for pattern matching, removing stop words & special symbols, converting the data to uppercase for easier comparison, replacing abbreviations of organizations with their original names and data management. The databases were also created in MySQL by joining tables based on a primary key to allow easier search and query for the client. Handling large amounts of data, importing and merging multiple files into the database as tables was not only challenging but also difficult given the system's capabilities. Our team extracted a list of all the attributes in our data and used them to link the tables within the database. Importing the data also required the data file format to be converted from excel(.xlsx) to a comma separated values format (.csv). As MySQL server only imports data from csv files, this conversion had to be done for every file before it could be used.

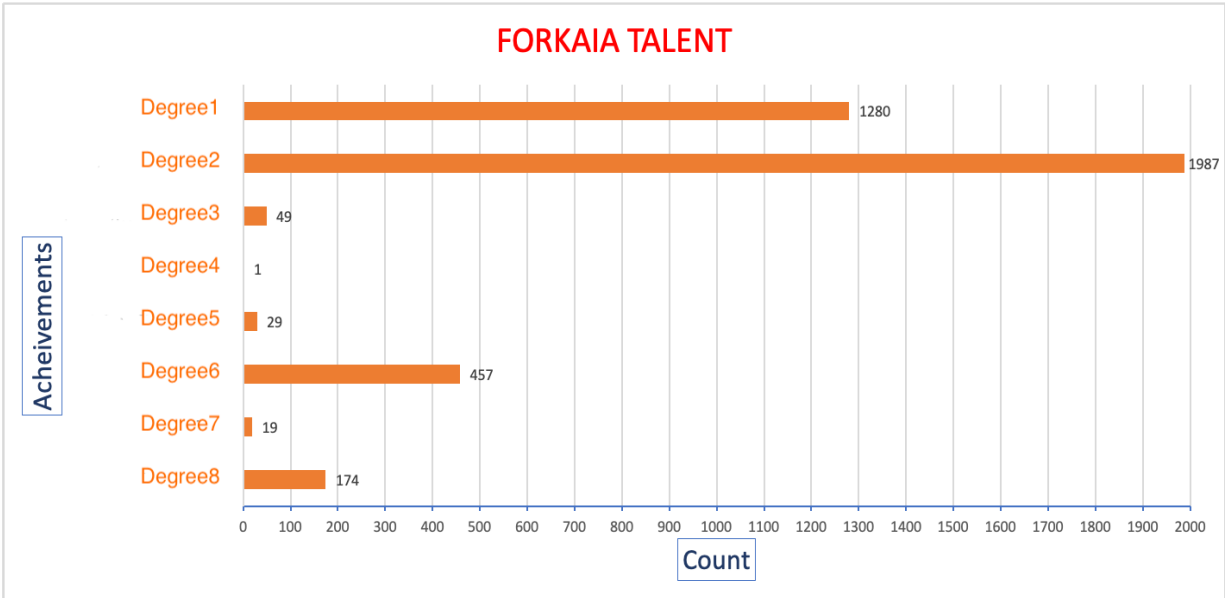


Fig 1: Distribution of Degree's earned by Company's talents

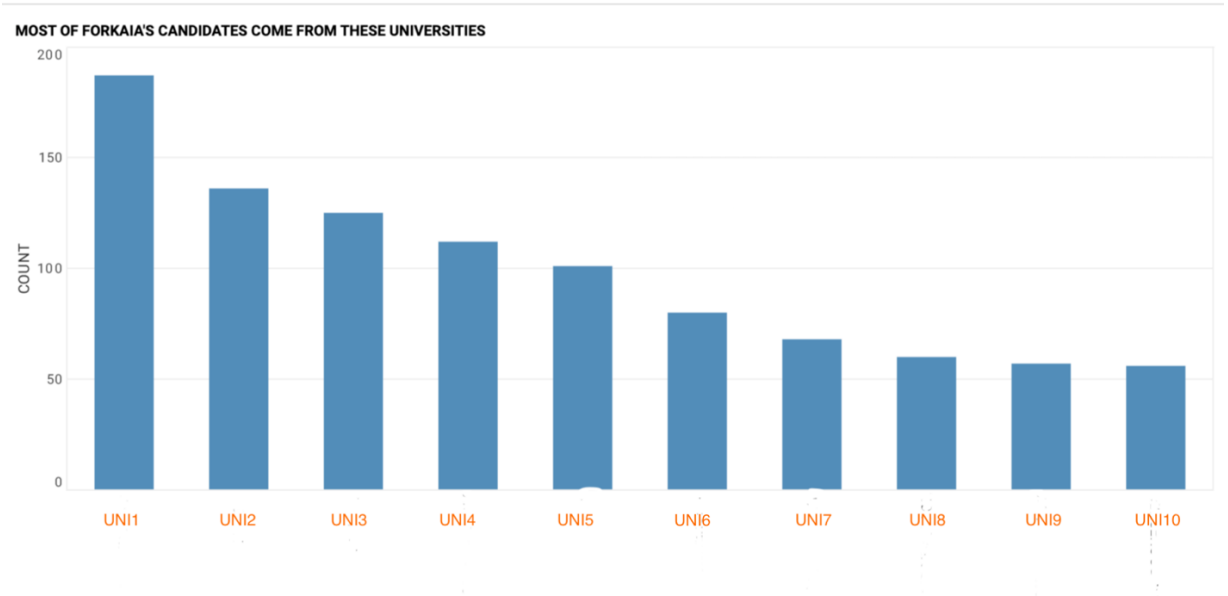


Fig 2: Showcasing backgrounds of the company's talents

The labels have been replaced with variables to protect the confidentiality of the company's data. For one of the tasks, the data was given as a preference for each candidate to work in certain teams for certain startups, the result consisted of each Startup team broken down by the number of people's requests to work on those teams. This information was extracted through NLP techniques, text-extraction, plots in Python (matplotlib) and pattern matching from the applications of candidates. I automated the entire process for future use of the company, such that, a single executable file of code will take the new raw data, perform pre-processing and generate analytical results and represent them visually. These figures show a few results amongst many insights I gathered from the company's application data files, to help understand the key demographic that the company appeals to and who appeal to the company as well.

6. **Learning experience:**

The internship was a great opportunity that enhanced my learning and allowed me a chance to practically use the skills I have learned during the course of my academics. Not only this, it also gave me an insight into entrepreneurship, business decisions and the power of teamwork.
The internship gave me a chance to be a part of multiple teams and work with my colleagues to find innovative solutions to the problems. It also gave me a chance to work on some individual tasks directly with the project manager which taught me responsibility and accountability. It allowed me to learn skills that are useful for a data scientist and analyst. Working at FORKAIA, I realized the importance of the balance between flexibility and structure and how the work environment can positively impact the productivity of the employees.

In CSE572 Data Mining, I learned to extract useful information (knowledge) from data through data mining techniques and present the results. I also learned ML techniques for training a model to be used for predictive analysis. The knowledge gained from this course was a fundamental pre-requisite that I took to my role as a Data Science & Analyst Intern at FORKAIA. In CSE578 Data Visualization, I learned how to create user-friendly visualizations, which I incorporated in my results during most of the tasks at my internship. In CSE 511 Data Processing at Scale, I learned database management and how to handle big data, which was a crucial part of my database part of the project.

As a part of this internship, I gained many insights, such as: time management, deadline management, importance of team work and the importance of a positive work culture. The positive work environment motivated me to perform my best, helped me enhance my knowledge and skills and working as a team made me realize the power of uplifting each other. The internship also gave me experience of working in the industry and reassured me of my decision to pursue my professional goal of working as a Data Scientist. I developed many technical and soft skills during this internship. The technical skills include text extraction techniques, pattern matching, analysis techniques and using platforms like Tableau for creating powerful and insightful visualizations. The soft skills include task management and organization, punctuality, maintaining communication with the team and the supervisor, and meticulousness in my work.

7. **References:**

[1] https://www.forkaia.com/ecosystem
[2] https://www.forkaia.com/combine