

**DSCI 6002\_3 Group 2 Final Project**

**TECHNICAL REPORT**

****

**FALL 22**

|  |  |
| --- | --- |
| **CONTENTS** | **1** |
|  |

**Project Name** **2**

**Executive Summary** **2**

**Technical Report** **3**

**Highlights of Project** **3**

**Submitted on:** **3**

**Abstract** **4**

**Methodology** **5**

**Results Section** **8**

**Conclusion** **8**

|  |  |
| --- | --- |
| **Current Real-World Job Skill Demand Vs Near-** | **2** |
|  |
| **Future Supplies From University Graduates** |  |

**Executive Summary**

**Have you ever heard students said that after investing years of precious time, money, and efforts in school, what they learned are not what industries are looking for? Our team is interested in finding out what are the real-world job skill demands today though job board postings, and how well students who about to graduate or to attend schools for a particular discipline what they learned, the supplies, matches with the demands.**

****

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**3**

**Technical Report**

****

**Title of the project**

**Current Real-World Job Skill Demand Vs Near-Future Supplies From University Graduates**

**Highlights of the Project**

**Our project provides students with the scalability of the skills they learned to become a data scientist in today's job market, as well as additional insights into which courses to take to complete the skillset that companies are looking for. Because the data collection involved new data from the past three weeks or less, we can confidently predict that the data will remain the same for at least the next four years and users will receive an accurate result.**

**Submitted on:**

**December 08 2022**

**4**

**Abstract**

**There are often mismatches of real-world job skill demands and the skills acquired by student through their curriculum at the school as not all universities’ programs are created equal and fast paced and frequent changes in tech world. In the first phase of the project, our team will zoom in to Data Science discipline to reveal the insights of job skill demands, and the scalability of the relevant skills a student acquired and learned or will learn in school, how well it matches the real-world job skill demands gathered from today’s job board. In the second phase our team would like to recommend additional skill to acquire to increase the chances of matching with the demand.**

**In our project we are using CRISP-DM and Agile SDLC methodologies to build a concrete and simple user interface from raw data available on Job portals. Our main business question is whether a student is equipped with the skillset required by companies in the present timeline. For the data to be up-to-date and results to be accurate we have collected data of job postings from multiple job portals and created databases to maintain and manipulate the data using SQL. As for the data preparation and visualization we used Jupyter Notebook extensively, with most of the python libraries to give a precise understanding on the job market. After careful data evaluations we deployed the user interface model onto Microsoft excel.**

**5**

**Introduction**

**Job skills are the abilities you use to complete your work, ranging from technical skills such as programming to workplace skills such as communication skills, leadership skills and more. A student is expected to have all of the necessary skillsets by the end of their course in order to be job-ready. A data scientist must have two types of skills: hard skills or technical skills and soft skills. The former is teachable and quantifiable, whereas the latter is innate and learned through experience. And our project’s main objective is to provide quantitative real-time results of skills acquired by students as well as additional insights into what supplementary skills one can learn. Because each university has its own set of curriculums, all students are not equally equipped with the same skills. On the other hand, skill sets required by industries are not the same, as they differ even by a minuscule amount for each job posting. Our team believes there is a need to understand the supply-demand gap and to construct an information bridge that benefits both students and employers. This requires a user-interface application that provides real-time results.**

**Also, every data science project has a business understanding and business questions at its core. Our main business questions are firstly what are the top skills required in the field of data science? Also, ow well one’s skills matching with the real-world? And the lastly when one skill is required, what’s the possibility of the other skill also required? We have a build the datasets and conducted analysis to answer all these questions.**

**We have used Agile and CRISP-DM methodologies over the course of the project which involved all the concepts like data collection, data pre-processing, data preparation, data modeling, data evaluation and model deployment. Because the job portals did not allow web scraping, our team manually collected 200 job postings from them. We used SQL and Python for data pre-processing till data evaluation.**

**Methodology**

**AGILE SDLC METHODOLOGY**

**The raw data collected by team members are notepad text files from multiple job portals namely LinkedIn, Glassdoor, Monster.com, Indeed which had different patterns of job posting. The data was formatted to usable SQL database tables with corresponding codes were developed in reference to the formats of each job portal. By following Agile methodology, we developed two iterations one with 40 jobs, after successful development we increased the job count to 165 in the next iteration. All the text files are imported to tables using SQL to make it look like simple and readable databases**

**6**

**CRISP-DM METHODOLOGY**

**As mentioned earlier, our team used CRISP-DM methodology for the analysis and preparation of the project. After collection of data from job portals in the form of notepad text files, the data was imported in JSON format with Jupyter notebook and deep analysis done to extract all the skills as part of qualifications or descriptions of all job posting thereby documenting to 90 hard skills and 12 soft skills. The data is then sent for pre-processing and multiple stages of data preparation to make it analysis and modeling ready.**

**Data Pre-Processing**

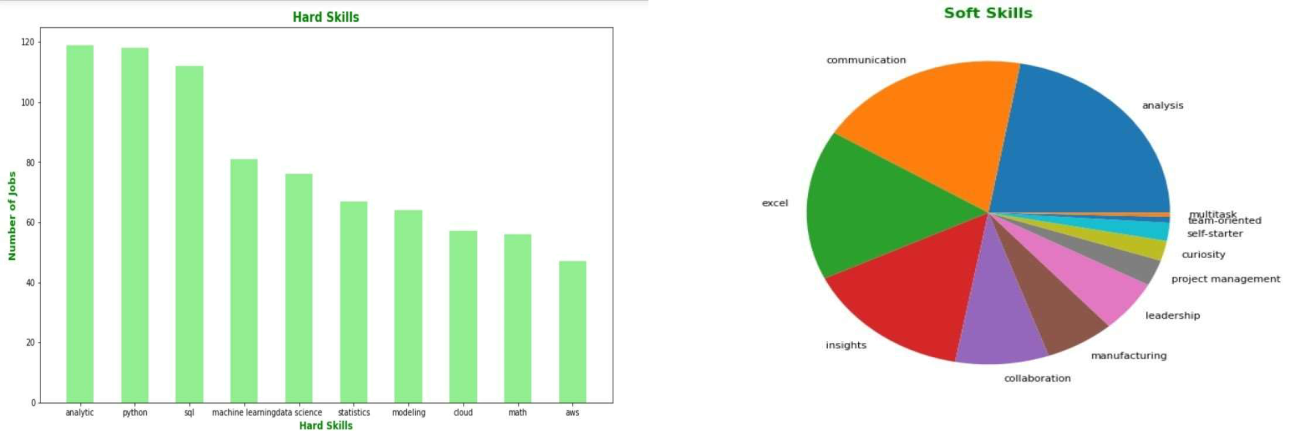
**As data collected from various job portals had dissimilar formats, and overwhelming content about company, job descriptions and more which do not account for any quantitative modeling needs to be removed, it is required to be pre-processed to check whether it is readable and remove all the error files which had defects. The text files are passed through various code to check for error before starting data preparation. 165 healthy files are used to build datasets out of 200 files.**

**Data Preparation**

**In the first stage of data preparation, text files were loaded and used to build a “Skills” master table with the help of tools such as SQL and Python. The Skills table consists of job count for a particular skill, job description and alternate skills needed for a particular skill given by skill Id, which gives very important data of job count vs skills. In the next phases, a “Jobs” table has been created with SQL which tabulated Job title, Job Id, company name, date posted, skills, URL and the source of and updated skill job count column in addition to updating skills jobs cross reference table “SkillsJobs”. Furthermore, job skills pairs well prepared to make the data feasible for data modeling. These tables have question first and second business question.**

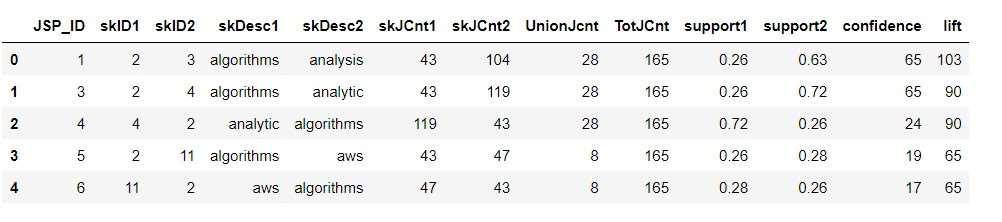
**Data Modeling**

**As a part of modeling data, listing out number of skills was the first task. 90 hard skills and 12 soft skills were identified with the data curated. Analysis of those skills are given by figures below as a picture speaks a lot.**

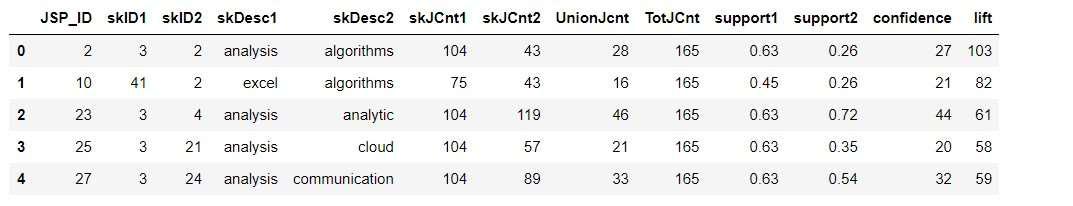
****

**7**

**To answer the third question which is about co-relation of two skills, association rules with Apriori has been used as to calculate probability of skill requirement by a company with two skills in combination. Two tables were modelled for the analysis of skill pair, one using only hard skills and other using soft skills in relevance with soft and hard skills. The glimpse of the two table with association rules with parameters support, confidence and lift is included in the below figures**

****

**Hard skills Association table**

****

**Soft skills Association table**

**Data Evaluation**

**The outcomes of the project are below listed in points**

**90 Hard skills and 12 soft skills are required in the field of data science across various job title such “Data Science”, “Data Analyst”, “Data Engineer”, “Artificial Engineer”, “Machine learning engineer” and more.**

**Analytics tops the charts in the Hard Skills as most of the companies and job listings require this skill. It is closely followed by Python SQL, other top skills include "Machine learning”, “Data science”, “Statistics”, “Modeling”, “Cloud”, “Math” and AWS.**

**Analysis skill tops the chart of Soft skills closely following by communication skills and excel expertise.**

**As for the association rules more than 500 combinations are deduced based on Apriori calculations**

**8**

**Model Deployment**

**A user interface skill evaluation tool has been developed and delivered through excel, which list out all relevant job with the skill student possess including the job location, company details, date posted and probability of satisfying the job requirement with skill acquired.**

**Results**

**In contrast to the most believed statement that Python, SQL and R are top skills required for a data scientist our project shows “Analytics” is the top hard skill required in the field of Data Science, followed by Python and SQL. The similar trend with soft skills reveals “Analysis” is the top skill required surpassing communication and excel skills.**

**Conclusion**

**To conclude, field of data science required many technical skills in reference to the job title, company and also paves a road to multiple job options with the same skills one possesses.**