LinkedIn Job Listings Analysis: Discovering Data Science Job Market Trends

Aishwarya Budhkar
Indiana University
Indiana, United States

Jagriti Kumari
Indiana University
Indiana, United States

***Abstract*—**

In this paper, we analyzed the current trends in data science job market and proposed a content-based job recommendation system. We also mined association rules between skills to understand the combination of skills required for a successful applicant. Furthermore, we worked on classifying job descriptions in different job roles like data scientist, data analyst, machine learning engineer, principal researcher, big data engineer, statistician and applied scientist to understand the similarities and differences in these job roles. Current data science job market is of dynamic and competitive nature and analyzing the recent trends is important for candidates as well as recruiters and hiring managers to find the right jobs and candidates respectively. Recent LinkedIn job listings data was web scraped using the linkedin-jobs-scraper library, licensed by MIT for this study. Exploratory data analysis is done to draw some insights about the education requirement, majors, most listed skills and domains.

***Keywords*—**

Web mining, Recommender systems, Clustering algorithms, Classification algorithms, Association rules, Machine learning, Boosting, Neural networks, Decision trees, Nearest neighbor searches.

I. INTRODUCTION

In today's job market, there are thousands of job postings at job sites everyday which makes it taxing for the job seekers to apply to the jobs relevant to him/her. This may result in the job seeker missing out applying to companies which are suitable for him/her that is where they have a good chance of getting selected as well as that matches their interest. Also, it has become challenging for recruiters and hiring managers to find the candidates suitable for a role they are hiring for due to several job seekers applying to the same role.

Hence, we worked on analyzing the latest trends in data science job market and build a tool that will help the candidates find the roles that is suitable for the candidate and of interest to him/her quickly. We have built a job recommendation system based on similarity in job descriptions posted on LinkedIn. We also used Apriori algorithm to find the association rules between the skills based on the listings on LinkedIn to understand what job skills occur together frequently. Moreover, we implemented a system that would classify a job description into different job roles of data scientist, data analyst, machine learning engineer, principal researcher, big data engineer, statistician and applied scientist to get more information about the type of job from the description. This will not only help the job seekers to apply to the different jobs at hand, but also given the skills, education and domain of his major, a job seeker may be directed to apply to the job of his interest and where he has a good chance of being selected and performing well.

The rest of the paper is organized as follows: Section II describes the motivation for the project; Section III describes the related work; Section IV explains about the dataset used; Section V talks about the exploratory data analysis; Section VI explains the design diagram; Section VII describes the actuation and remediation; Section VIII describes the pipelines and results obtained; Section IX states the conclusion derived from the work; Section X is about the limitations and possible improvements.

II. MOTIVATION

LinkedIn already has an abundance of job advertisements in its database, and hundreds of jobs are posted each day. With so many jobs, it has become difficult to find the actively hiring roles that are most relevant jobs for the candidate.

A job seeker tends to apply to the jobs that are most coveted. For example, a recent college graduate with major in MS in Data Science, may look for jobs that are titled 'Data Scientist' and may skip looking at the job description of a 'Research Analyst' that may have posted the roles suitable for a data scientist. For this reason, we have built a job recommendation system using cosine similarity to address the issue.

Moreover, potential candidates face many rejections from the companies that require a combination of skills. This may be because a person cannot possibly learn all the skills in the world. Understanding the combination of skills that are listed frequently for a job type, that the candidate is interested in, can help them do targeted preparation to increase his or her chance of getting selected at the company.

For universities, understanding the latest trends in the data science job market is important so they can tailor their curriculum to teach industry-relevant skills

III. Related work

There is a boom in the data scientist job market requirement recently. Some of the top emerging jobs are machine learning engineer, big data engineer and data scientist [1]. For companies it is required to hire right candidates with required skills and train employees so they can update their knowledge to learn new skills required to perform better at job. Hence, it is important to understand upcoming technical skills that will boost the business and provide training to employees [1]. Textual analysis of Glassdoor jobs has revealed, the rise in data scientists' jobs. sql and python are most listed skill requirements [2].

Github and StackOverflow analysis as well as job portal analysis reveals a need for a framework to detect the hard and soft skills which would be helpful for job seekers, hiring managers and HR [3]. Only text-based analysis does not reveal the importance of skills as it only checks the occurrence rather considering combination of skills in related roles gives better performance in job recommendation [4].

There is a need for educational institutions to adapt their curriculum to teach latest in-demand skills in industry to bridge the gap between industry-relevant skills and university courses [5]. Educational institutions play an important role in connecting the needs of industry and research. Soft skills are very important in industry along with technical skills but there exists less representation of soft skills in technical curriculums. There is a need to understand importance of soft skills and include in university curriculum [5].

The job advertisements on job search websites consists of structured data like job titles and unstructured data like the location of the company on the map [6]. The main attributes of a job posting are company name, job title, description of the job, date, seniority level and job requirements. Out of these, job title and description are the main features. The title as the name suggests is the name of the position that a candidate is looking for, and the description contains information like skills, qualifications and brief information about the job [6]. The job recommendation model aims at extracting information from the job posting using machine learning clustering methods. The jobs are divided into clusters based on features and similar jobs based on job description are recommended to the users. [6].

Equal emphasis on soft skills and technical skills is seen in the posted listings. Leaving 5% internships and 5% senior roles most listings seem to be for all levels. As the market is rapidly evolving, the recommendations are required to be updated constantly to stay relevant [8]. With big data tools, dynamic fetching and updation of data, a recommendation system can be developed for job seekers and recruiters [8].

IV. Datasets

We web-scraped LinkedIn data using linkedin-jobs-scraper [11] 1.8.4 licensed by MIT. We collected data for job titles Data scientist, data analyst, applied scientist, statistician, principal researcher, machine learning engineer, big data engineer. The size of data is 32MB with 6000 listings scraped. The data is static and collected once.

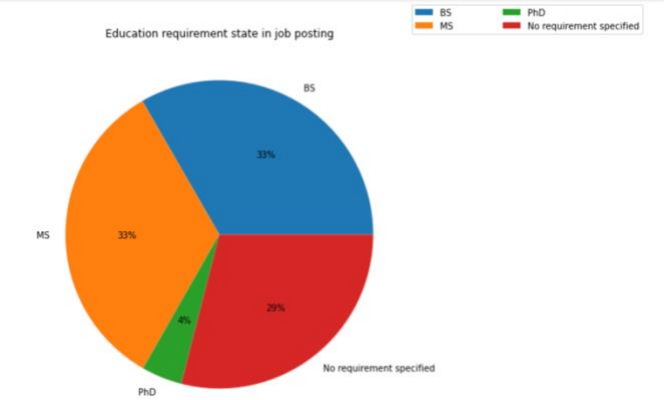
Following fields were available in the data:

Column Name	DataType	Example Values
JOB ID	INTEGER	2801690476
TITLE	STRING	Data Analyst
COMPANY	STRING	Webmasters4SEO
DATE	STRING	2021-11-20
LOCATION	STRING	United States
PLACE	STRING	Dallas, TX
JOB FUNCTION	STRING	Information Technology
EMPLOYEMENT TYPE	STRING	Full-time
INDUSTRIES	STRING	Financial services
DESCRIPTION	STRING	Think big. Dare to …

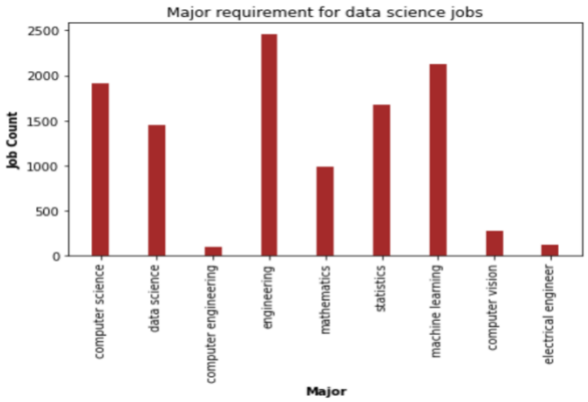
V. Exploratory Data Analysis

The analysis was done for entire data as well as for different job types to get job type specific results. Following are the results:

As seen in above Figure, for education, we saw that 33% percent mentioned bachelor's and 33% mentioned master's degree requirement while 29% did not mentioned any degree level requirement. For 4% jobs doctorate degree was listed.

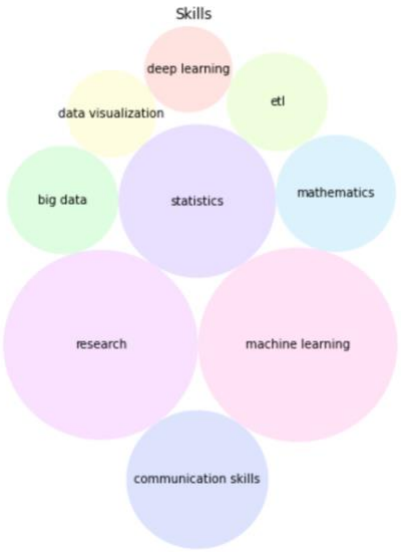


As seen in below figure, we see that engineering, machine learning, data science, computer science and statistics majors are desired.

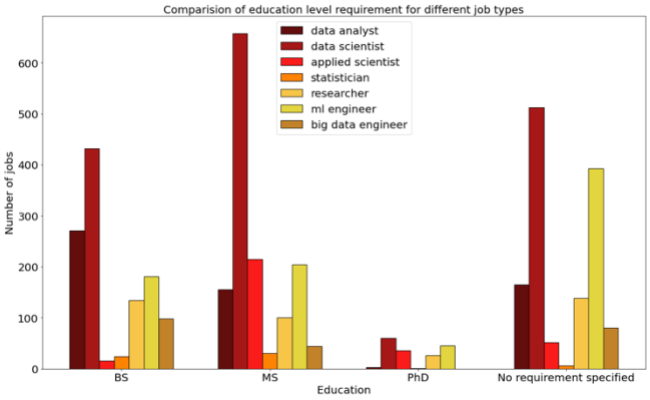


As seen from the below figure, among technical skills, python is the most desired skill followed by Aws, Scala, Spark, Java, Hadoop and Tableau

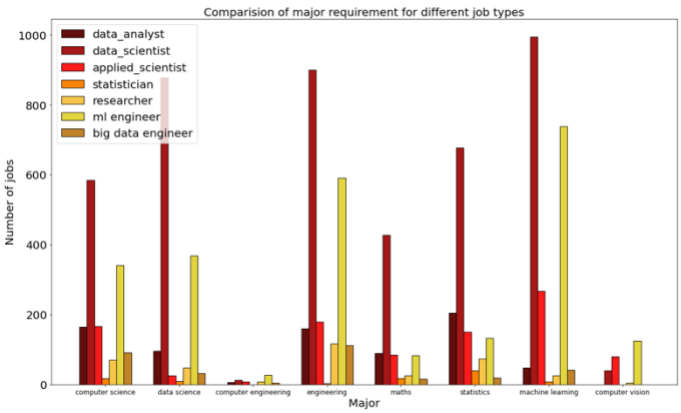
As seen below, Machine learning, Research, Statistics, Mathematics and communication domains are mentioned more often than others among the jobs in data



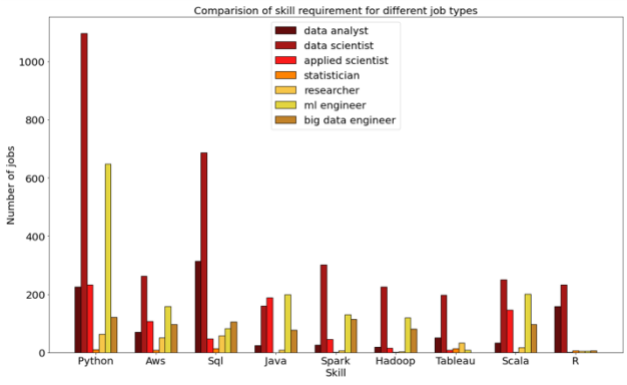
a.



From the data we infer that Bachelor requirement for Data analyst and big data engineer is more followed by Masters and PhD while requirement for Masters requirement is more for Data scientist and Applied scientist. Masters and Bachelors graduates are equally desired for Researcher, ML engineer and Statistician. Many job listings haven't specified any education level requirement.



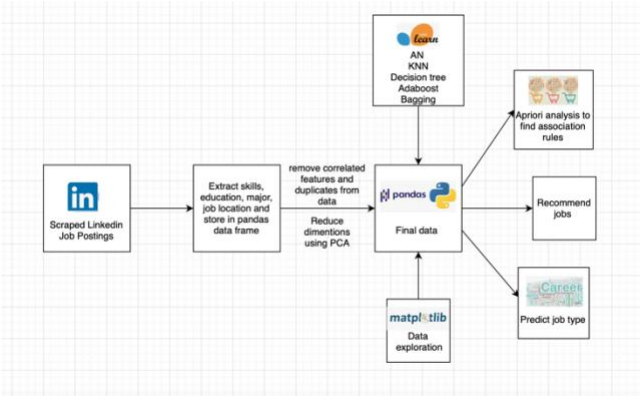
We see from previous bar chart, for ML and big data engineer machine learning and engineering majors are more important. For researcher and statistician, statistics, math and engineering majors are more desired. For applied scientists and data scientists, machine learning is most important followed by data science, statistics, engineering and computer science. For data analyst statistics and computer science are more desired.



As seen above, for data analyst sql is most important skill followed by python. For data scientist python is most desired followed by sql. For applied scientist, python, java, scala are most listed skills. For statistician tableau is most listed followed by sql and python. For researcher python, aws, sql are equally listed. For ml engineer and big data engineer python, java, scala, aws, spark, hadoop are listed with python prominent is the former followed by others while all equally listed in the later.

VI. Design diagram

Following design diagram describes the process flow. LinkedIn job postings were scraped using web scraper. First, data is cleaned by removing duplicated and null values. Next, we extract the desired information from the job description like skills, education, major and job location state. Attributes are discarded which are correlated to each other. Next, dimensionality reduction is done with PCA. Only one of correlated attributes is used for further analysis. We use the final cleaned and reduced data for skills association analysis and job type classification. ML models are trained to predict the job type using the description. Apriori association rule mining algorithm is used to find skills that occur together. Content based job recommendation system is developed to recommend jobs based on job description similarity between roles that the candidates is interested in.



VII. Actuation or remediation

1. Any description can be classified into the job type so that job seekers get better idea about their role and expectations from them.
2. Closely associated skills for a particular job type can be found out and job seekers interested in that type of job can be provided recommendation to learn the related skills to increase their chances of getting selected and performing well at the job.
3. Content based recommendation system will suggest jobs to candidates which will help them apply to job profiles that they are interested in rather than concentrating on job titles alone.

VIII. Pipeline And Results

The first step is data cleaning. Duplicates are removed from web-scraped data. Skills, education, major and Job location are extracted from job description. The extracted data had many inconsistent values. For example. Masters degree is mentioned as Master, Masters, Master's, MS, etc. Thus, all the values were converted to consistent representation. Pandas data frame and matplotlib was was for data exploration. To predict type of job based on job description, ML models from scikit learn were used. K-nearest neighbors, Artificial neural network, Decision tree, Adaboost and Bagging with base estimator decision tree models were used to train the model to use to features and classify the job description. To make the class distribution uniform among the records, Synthetic Minority Over-sampling Technique [12] is used. The data needed to be one-hot encoded, and features were lists. For example, list of skills. After one-hot-encoding the number of features became 102. To reduce dimensions, Principal component analysis is performed on data. Correlated attributed like numpy, pandas and sckit-learn and tensorflow, pytorch were converted to single feature. After dimensionality reduction, 40 features were chosen for training machine learning models. Data is split into train and test with 30% test data. Machine learning models were trained with Train data and accuracy and F1 score is calculated to find the best performing model. Following table shows the results:

Algorithm	Test accuracy	F1 score
KNN (5 neighbors)	0.844329	0.840934
Decision tree	0.891363	0.890001
Adaboost (base estimator decision tree, 5 estimators)	0.899896	0.899021
Bagging (base estimator decision tree, 5 estimators)	0.895525	0.894284
ANN (2 hidden layers 256 128 relu activation)	85.72%	

We obtained best performance with Adaboost with base decision tree and 5 estimators. The models were able to distinguish between the roles based on education, skills and major.

Apriori algorithm was used to find skills and domains that occur together in the description. We got the following results. We can recommend learnings to candidates based on the rules found to improve their chances of getting selected and performing well at the job.

	antecedents	consequents
0	(spark)	(python)
1	(java)	(python)
2	(sql)	(python)
3	(python)	(sql)
4	(scala)	(python)
5	(aws)	(python)

Lastly, we used similarity-based clustering for job recommendations. The descriptions are used to generate a word vector to find the similar descriptions and recommend to a job seeker job like a job type that he is interested in.

```
job_recommender('data scientist')
```

```
array(['research analyst', 'sr. data scientist', 'senior data scientist',
       'senior research analyst', 'data scientist', 'research assistant',
       'data scientist', 'data scientist', 'data scientist'], dtype=object)
```

IX. Conclusion

We could successfully use the LinkedIn data to generate models which can classify a job description in 7 types: Data scientist, data analyst, principal researcher, statistician, applied scientist, ml engineer and big data engineer. Also, we could find skills that occur together which can be used to recommend trainings to users. Finally, based on a description, a model was developed to cluster similar jobs to be recommended to job seekers. There are abundant data science job postings on LinkedIn and finding the relevant jobs and skills required for a job that he is interested in, quickly and easily will be a helpful tool for the job seeker. With more data behavior-based recommendation engine can be developed for user-centric recommendations.

X. Limitations and Future Work

By collecting more data like user reviews and other user behavioral features like job likes, comments, etc. we can develop a recommendation engine to recommend relevant job to different users. Also, we can recommend potential candidates to recruiters and hiring managers who have required skills and a good chance of performing well in the listed job. More widespread data like monthly or yearly can help us analyze trends in the job posting which can help us understand when the job market is most active in the year and accordingly tailor our recommendation to job seekers who are actively and passively looking for jobs.

Acknowledgment

We are thankful to the Professor Yuzhen Ye for her guidance and support throughout the course. We are also thankful to the graders for their help with the assignments and project.

References

1. S. Gottipati, K. J. Shim and S. Sahoo, "Glassdoor Job Description Analytics – Analyzing Data Science Professional Roles and Skills," 2021 IEEE Global Engineering Education Conference (EDUCON), 2021, pp. 1329-1336, doi: 10.1109/EDUCON46332.2021.9453931.
2. R. B. Mbah, M. Rege and B. Misra, "Discovering Job Market Trends with Text Analytics," 2017 International Conference on Information Technology (ICIT), 2017, pp. 137-142, doi: 10.1109/ICIT.2017.29.
3. M. Papoutsoglou, N. Mittas and L. Angelis, "Mining People Analytics from StackOverflow Job Advertisements," 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2017, pp. 108-115, doi: 10.1109/SEAA.2017.50.
4. Baoxu Shi, Jaewon Yang, Feng Guo, and Qi He. 2020. Salience and Market-aware Skill Extraction for Job Targeting. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 2871–2879. DOI:https://doi.org/10.1145/3394486.3403338
5. Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, James A. Evans, "Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy", Proceedings of the National Academy of Sciences Dec 2018, 115 (50) 12630-12637; DOI:10.1073/pnas.1804247115
6. D. Mhamdi, R. Moulouki, M.Y. El Ghoumari, M. Azzouazi, L. Moussaid, "Job Recommendation based on Job Profile Clustering and Job Seeker Behavior", Procedia Computer Science, Volume 175,2020, Pages 695-699, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.07.102.
7. N. A. Ismail and W. Z. Abidin, "Data Scientist Skills," IOSR Journal of Mobile Computing & Application (IOSR-JMCA), Volume 3, Issue 4 (Jul. – Aug. 2016), PP 52-61.
8. A. Shirani, "Identifying Data Science and Analytics Competencies Based on Industry Demand," Issues in Information Systems, Volume 17, Issue IV, pp. 137-144, 2016.
9. A. MOUMEN, E. H. BOUCHAMA and Y. EL BOUZEKRI EL IDIRISSI, "Data mining techniques for employability: Systematic literature review," *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 2020, pp. 1-5, doi: 10.1109/ICECOCS50124.2020.9314555.
10. I. Khaouja, I. Rahhal, M. Elouali, G. Mezzour, I. Kassou and K. M. Carley, "Analyzing the needs of the offshore sector in Morocco by mining job ads," 2018 IEEE Global Engineering Education Conference (EDUCON), 2018, pp. 1380-1388, doi: 10.1109/EDUCON.2018.8363390.
11. https://pypi.org/project/linkedin-jobs-scraper/
12. Chawla, N.V. Bowyer, K.W. Hall, L.O. Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 2002, 16, 321–357.