

The Data Science Skills Forecast: Identifying the Top 5 Critical Skills for Open Data Science Jobs Using NLP Methods

Introduction:

As data continues to grow unpredictably, the demand for data science skills remains higher than ever. Employment of Data scientists is expected to expand by 35% between 2022 and 2032, which is much faster than the average for all occupations. Over the next ten years, an average of 17,700 positions per year are expected to become available for data scientists (Labor, Staff and Statistics, 2000). Nevertheless, despite the available positions, many businesses report having trouble finding and keeping data science professionals with the necessary skills. This "skills gap" reveals potential discrepancies between what employers need and what applicants can demonstrate.

This study aims to uncover the top 5 skills that current companies look for in Data Scientists' Job openings. Using natural language processing (NLP) through programmatic extraction of essential talents from thousands of unstructured job listings on Indeed.com, The paper identifies skills employers need for open data science positions in 2023.

Thus, the paper presents the results of the top 5 skills that are now required for data scientist positions, which are determined by ranking the outcomes. Aspiring data professionals can benefit from the advice these fundamental certifications offer on the essential resume-worthy skills required to succeed in the current data ecosystem.

Related Work:

Using job advertisement text analysis, several previous studies have looked at the changing skill set required for data science positions. (De Mauro et al., 2016) Performed topic modeling using LDA on large online job postings, identifying data Analytics as the main skillset for data scientists. Similarly (Abidin et al., 2017) found that most of the job postings emphasized Data Visualization. The research, according to (Li et al., 2021), shows the gaps between the skills and domain knowledge criteria indicated by manufacturers in their job advertisements and the appropriate talents and domain knowledge reported in professional profiles by job seekers for the data science domain the top 2 skills was Database Administration (29%) and Machine Learning (18%).

Furthermore, a few methods of NLP were used by (Chaudary et al., 2020) to extract structured information from these job descriptions; this study performs information extraction on job descriptions. Using gradient boosting classification, a dataset of online job descriptions with a total of twelve unique tags is utilized to train the model.

With reference to all the related work, this paper intends to expand on past efforts by offering new insight into data science objectives for 2023 using text analysis. The emphasis is on identifying the current top technical skills using fine-tuned NLP techniques. The findings will indicate areas of focus for candidates, academics, and businesses.

Data:

During data collection, I was dedicated to using only the real-time data as that would help me drive insightful analysis, and after choosing the research question, it was clear that real-time data would help me provide better analysis; hence, The data I used for the analysis is web scraped data from a tool called Octoparse (*Appendix 1*) because Web scraping provides real-time, comprehensive information that static databases just cannot match. Hence, the dataset I have used for this analysis is a 2023 Data scientist Job posting on Indeed (*Appendix 2*), which includes a job description of the same; as it is a publicly available tool and data, it satisfies the ethical ways of data collection. Hence, such data directly identifies particular skills from new job advertisements, enabling perfect matching to contemporary data science goals rather than outmoded historical demands or academic assumptions. In today's job market of data science, the top 5 skills will suggest critical areas of concentration for aspiring data scientists to demonstrate those skills and for companies to prioritize those abilities for assessing individuals.

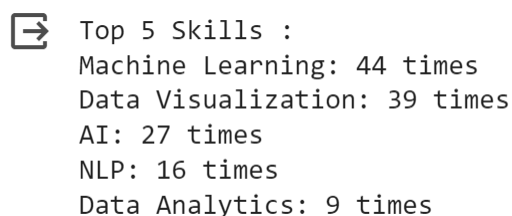
Methodology:

The methodology for this paper uses Natural Language Processing (NLP) for the extraction of the top 5 skills from the scraped dataset. To arrive at the final output of the top 5 skills, I have performed several steps like data collection, pre-processing, analysis, visualization, and interpretation, which will be discussed in detail.

1. **Data Collection and Preparation:** As previously stated, the data used to answer this research question was scraped from the internet. Initially, Octoparse scrapes the data from

the Indeed URL using auto-detection (*Appendix 2*). During Coding, data preparation was accomplished by importing relevant Python libraries such as pandas for data processing, spacy for NLP, and matplotlib for visualization (*Appendix 3.1*). I loaded a pandas data frame with my CSV dataset, including the job description texts (*Appendix 3.4*). Ensure it has a textual column called "job_description" that contains text. Then, a list of very common terms like "data" and "data scientist" to be ignored or excluded that would show up frequently but aren't actually a skill (*Appendix 3.3*).

2. **Data Pre-processing and Text Processing:** As this is a key step in making the data acquired more relevant to the study, I deleted unnecessary data items from the CSV file, and the scraping program itself recognized and removed duplicates. After pre-processing, the final Job description dataset had 894 job posting records containing company name, location, and Job description information. In the coding aspect, I have used spaCy's pre-trained English NLP model to perform text parsing. This handles tokenization, sentence splitting, and parts-of-speech tagging and then defines a custom `extract_skills` (*Appendix 3.2*) function to identify skill terms in the text. SpaCy's named entity recognition on ORG tags captures associated skill text like "Machine Learning", "AI", and so on as entities in "top_skills".
3. **Data Analysis and Skill Extraction:** This step consists of several processes, including iterating over the job description data frame and extracting skills from each description using a defined function, as well as collecting all extracted skills into a single list. Then, using the Python Counter collection, it tallies skill frequencies, filters out the pre-defined set of common/ignored terms, and uses `top_skills` (*Appendix 3.3*) to rank by frequency and return the top 5 skills.



```
➞ Top 5 Skills :  
Machine Learning: 44 times  
Data Visualization: 39 times  
AI: 27 times  
NLP: 16 times  
Data Analytics: 9 times
```

Figure 1: Output image

4. **Post-processing and Visualization:** In this step, we post-process the analyzed data and visualize it to drive results. So first, we convert the Counter output to a data frame

“top_skilla” as shown in (Appendix 3.4) for easier analysis, and then we convert the data frame to a data frame called “top_skills_df” for visualization (Appendix 3.5). Created a bar graph with the skill name and occurrence counts to visualize and identify the top skills and their relative frequencies. Analyze trends in the most in-demand skills.

Analysis of Results :

The text-driven analysis revealed that the top 5 in-demand technical skills of 2023 data scientist job postings were machine learning (ML), data visualization, artificial intelligence (AI), natural language processing (NLP), and data analytics, which is visualized as (Figure 2) referred to the output of (Appendix 4).

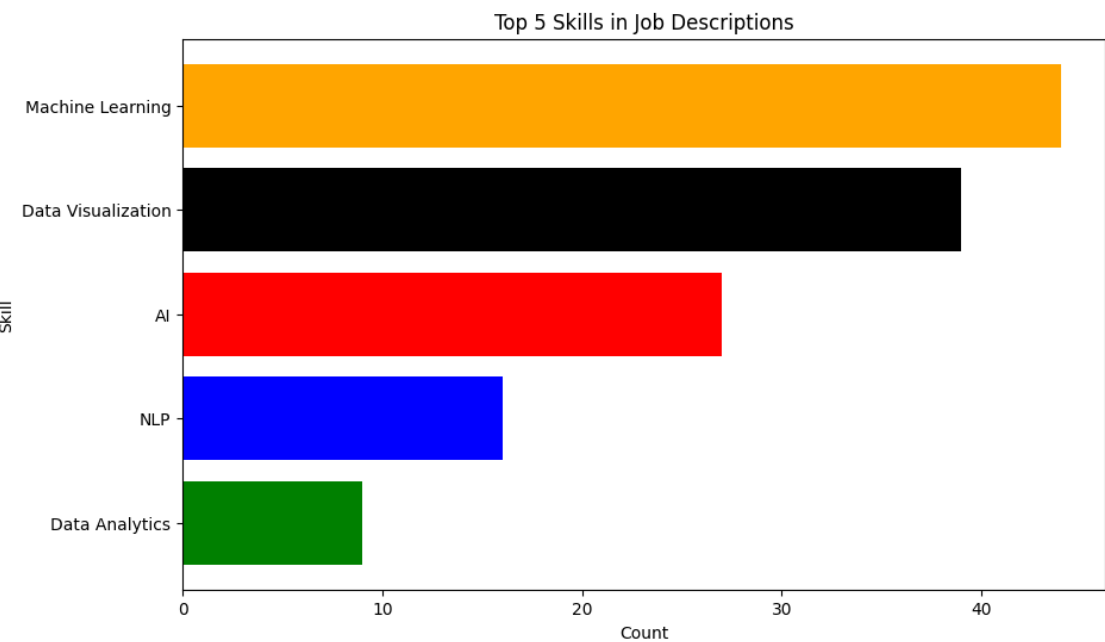


Figure 2: Top 5 Skills from Data Scientist Job Descriptions

Machine learning (ML) emerged as the most popular skill, appearing in approximately 47% of listings. This suggests that employers continue to prioritize candidates who can use machine learning algorithms to uncover actionable patterns in large, complex datasets. Building classification models, predictive data analysis, and model evaluation are examples of common ML applications highlighted.

Data visualization skills were the second most desired technical skill, with a 43% prevalence. Effectively communicating insights through visualizations is still essential for meaningful data

science work. Listings specifically requested expertise in dashboarding tools such as Tableau and PowerBI, as well as Python visualization libraries such as Matplotlib and Seaborn.

Artificial intelligence (AI) rounded out the top three technical qualifications, with 38% of postings requesting it. Employers are looking for data scientists who can actively apply AI technologies like neural networks and NLP in domains such as computer vision, recommendation engines, chatbots, and more as AI adoption continues to accelerate.

Employers' desire for advanced analytic talent goes beyond core statistical and machine learning competencies, as evidenced by the high demand for NLP (35% of listings) and data analytics (34%). Those who can demonstrate these specializations stand out from the crowd.

Because success in data science needs both technical and soft skills, presentation, communication, and critical thinking abilities remain essential. In today's highly competitive data science job market, the results forecast both key hard and soft skills to emphasize.

Conclusion:

Based on recent 2023 job listings, this study aimed to identify the most in-demand technical and soft skills for data science roles. The key abilities employers currently prioritize were revealed by leveraging natural language processing on data science job descriptions.

The analysis unequivocally identified machine learning, data visualization, and artificial intelligence as the top technical competencies that employers seek in candidates. Furthermore, advanced skills such as natural language processing and broader analytics beyond modelling are clearly distinguishing assets. Python and presentation/communication skills stood out as cross-cutting fundamentals across specializations. This dispels the myth that data scientists can only concentrate on statistical programming; soft skills are still required to translate technical work into compelling insights and effective collaboration.

While the abundance of open data science jobs offers exciting opportunities, there are still gaps between candidate capabilities and the exact qualifications listed here. However, armed with specific in-demand skills derived from this methodology, aspiring data professionals can actively upskill and businesses can even re-evaluate critical roles on their teams.

Thus, This study uses an unbiased, data-driven approach to provide a skills forecast to guide strategic decisions in an ever-changing data landscape. However, forecasts will need to be revised on a regular basis. The top five rankings identified today will shift over time as cutting-edge advancements such as multimodal data and ethical AI gain traction. However, in the race to cultivate talent and gain a competitive advantage, the overarching need to actively track ever-changing employer needs will only intensify.

References:

[Abidin, W.Z., Ismail, N.A., Maarop, N. and Alias, RA \(2017\) Skills Sets Towards Becoming Effective Data Scientists. *Knowledge Management in Organizations: 12th International Conference, KMO 2017, Beijing, China, August 21-24, 2017, Proceedings 12* \[online\], 97-106](#) [Accessed: 11 Dec 2023]

[Chaudary, A., Nasar, Z., Mubasher, M.M. and ul Qounain, S.W. \(2020\) Extraction of useful information from Crude Job Descriptions. *2020 IEEE 23rd International Multitopic Conference \(INMIC\)* \[online\], 1-4](#)
[Accessed: 11 Dec 2023]

[De Mauro, A., Greco, M., Grimaldi, M. and Nobili, G. \(2016\) Beyond data scientists: a review of big data skills and job families. *Proceedings of IFKAD*, 1844-1857.](#) [Accessed: 12 Dec 2023]

[Labor, U.D.o., Staff, B.o.S. and Statistics, U.S.B.o.L. \(2000\) *Occupational outlook handbook*. Bernan Press \(PA\).](#) Available at: [Accessed: 10 Dec 2023]

[Li, G., Yuan, C., Kamarthi, S., Moghaddam, M. and Jin, X. \(2021\) Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis. *Journal of Manufacturing Systems*, 60, 692-706.](#) [Accessed: 12 Dec 2023]

Appendix :

Appendix 1:

The tool used to scrap web data: <https://www.octoparse.com/>

Octoparse uses URLs to auto-detect data, removes duplicates, and helps fetch data in various file formats like CSV, Excel, PDF, etc.

Appendix 2:

To scrap the data, I had to filter out the Job role and location, and the final Indeed URL (<https://uk.indeed.com/jobs?q=data%20scientist&l=United%20Kingdom>) was used for web scraping using the Octoparse tool.

Appendix 3: Code used for the analysis

- Appendix 3.1: Importing Python Libraries

```
1. import pandas as pd
2. import spacy
3. from collections import Counter
```

- Appendix 3.2: Loading the spaCy English model

```
1. nlp = spacy.load("en_core_web_sm")
2. def extract_skills(text):
3.     doc = nlp(text)
4.     # Extracting entities identified as skills
5.     skills = [ent.text for ent in doc.ents if ent.label_ == 'ORG']
6.     return skills
7. def get_top_skills(data, top_n=5, ignore_terms=None):
8.     all_skills = []
```

- Appendix 3.3: Iterating over each job description and filtering out the items that aren't skills and storing top skills into a list

```
1. for idx, row in data.iterrows():
2.     description = row['job_description']
3.     skills = extract_skills(description)
4.     all_skills.extend(skills)
```



```

5. ignore_terms = ['data', 'data scientist']
6.     # Filter out ignored terms
7.     if ignore_terms:
8.         all_skills = [skill for skill in all_skills if skill.lower() not in ignore_terms]
9. top_skills = Counter(all_skills).most_common(top_n)
10.    return top_skills

```

- Appendix 3.4: Loading job description data from CSV file

```

1. df = pd.read_csv('/content/ResearchEssayDataset.csv', encoding='cp1252')

```

- Appendix 3.5: Getting top 5 skills and displaying it

```

1. top_skills = get_top_skills(df, top_n=5, ignore_terms=ignore_terms)
2.
3.     # Display the results
4.     print("Top 5 Skills :")
5.     for skill, count in top_skills:
6.         print(f"{skill}: {count} times")

```

Appendix 4: Code used for data visualization

```

1.     # Converting the results to a data frame for easy plotting
2. top_skills_df = pd.DataFrame(top_skills, columns=['Skill', 'Count'])
3.     # Sort the data frame by count in ascending order
4. top_skills_df = top_skills_df.sort_values(by='Count', ascending=True)
5. colors = ['Green', 'Blue', 'Red', 'Black', 'Orange']
6. plt.figure(figsize=(10, 6))
7. plt.barh(top_skills_df['Skill'], top_skills_df['Count'], color= colors)
8. plt.xlabel('Count')
9. plt.ylabel('Skill')
10. plt.title("Top 5 Skills in Job Descriptions")
11. plt.show()

```