Career Advancement Tool (CAT)

Yogev Namir - 318880754 , Yonatan Sabag - 316277219, Ori Barazani - 316137371

Github repo - https://github.com/yogev-namir/Linkedin-Career-Advancement-Tool-CAT

Google Drive data folder - https://drive.google.com/drive/folders/1srN0luAgNj1ycUqMDPhjQo5GTVZv4DkO

Project Introduction

The Career Advancement Tool (CAT), a new feature on LinkedIn, enhances job seekers' marketability by addressing the gap between the skills currently listed in user profiles and those demanded by the evolving job market. CAT uses machine learning and large language models to analyze and suggest improvements to resumes, ensuring they meet industry standards. It also helps users acquire or improve skills to match these standards. CAT is necessary due to the dynamic nature of the job market, where traditional job seeking and skill presentation methods are increasingly inadequate. It offers a personalized, data-driven approach to professional development, not only refining resumes but also identifying skill gaps and suggesting learning pathways through LinkedIn Learning. This comprehensive strategy aims to ensure users' profiles surpass industry standards, boosting their job match potential and supporting continuous professional growth.

Data Collection and Integration

For the **Data Collection and Integration** section of the **Career Advancement Tool (CAT)** project, we have employed a two-fold strategy: utilizing existing datasets from LinkedIn and enhancing these with additional data collected through web scraping.

Original Datasets Utilized: The core of our data comes from LinkedIn's profiles dataset, which is rich in details about users' professional backgrounds, educational histories and experiences. This dataset is pivotal for CAT's functionality, allowing us to:

- Conduct professional development analysis by examining employment history, and educational achievements.
- Provide customized recommendations for aligning users' resumes with industry benchmarks and identifying areas for enhancement.
- Offer personalized skill advancement guidance and recommend relevant LinkedIn Learning opportunities.

Additional Data Collection and Integration: Our team has enhanced the Career Advancement Tool (CAT) by enriching the LinkedIn dataset with current job market data and course information from LinkedIn Learning, reflecting the latest professional trends. We achieved this through web scraping for new courses and integrating a comprehensive job skills dataset from Kaggle, dated 2024. We focused on condensing over 10,000 skills down to the 100 most critical, aligning with job market demands, which are now detailed in our reports. Additionally, we developed a neural network to predict skills from job titles and companies. By leveraging these updates, CAT now offers more precise and up-to-date career development support, particularly for users interested in the growing data analysis sector.

Items and Enrichment Size: In our project, an "item" is defined based on the context of the dataset it belongs to:

- LinkedIn Profiles: An item is a user's profile, including professional experience(past jobs), skills (predicter via NN), education (degrees), and current job (job position name).
- LinkedIn Learning: An item is a course, complete with name, rating, and enrollment numbers. We used a scrapped data from Kaggle with 12,218 items.
- Job postings: An item is a job title, job posting publisher name and required skills. We managed to scrape 974 items from Linkedin Learning website.

For conclusion, we enriched our original given Profiles dataset with the help of 13,192 new items (the Profiles size remained the same but the new items helped us enriched it).
Alpha = 13,192 / 2.5M = 0.528%

Data Analysis

In selecting key features for our analysis, we prioritized data quality and relevance, guided by insights gained from extensive engagement with the dataset throughout the semester. Our focus was on features with minimal missing values and significant relevance to the Career Advancement Tool (CAT) objectives. Profiles dataset features utilized are:

- **Education**: We quantified the number of distinct and valid degrees listed in a profile, emphasizing the depth of educational background.
- **Experience**: The count of unique and verified previous job roles a profile has held was determined, highlighting the breadth of professional experience.
- **Position**: The current job title of the profile was meticulously cleaned of any extraneous information before being processed. This title then underwent tokenization, ensuring it was optimally prepared for our analytical models.

AI Methodologies

In the development of the Career Advancement Tool (CAT), our project utilized a mix of advanced AI methodologies, algorithms, and thorough evaluations to ensure the effectiveness and accuracy of our recommendations. Here's a detailed look at the AI strategies we implemented:

DistilBert for Tokenization: For processing job titles, we utilized DistilBert, a streamlined version of the Bert model. DistilBert efficiently converts job titles and publishers (company name) into 768-dimensional vectors, facilitating a deeper and more nuanced analysis. This tokenization process is critical for matching users with relevant skills and job profiles, enhancing the personalized recommendations provided by CAT.

Neural Network for Skill Prediction: We employed a neural network to predict the set of skills associated with given job titles and publishers (company name). This deep learning model allowed us to understand and decode the complex relationships between job titles and companies names and their associated skills. By training the neural network with a large dataset of job titles and skills, we

ensured that CAT could accurately suggest skill enhancements for users' profiles, aligning them with current industry demands.

Bucketed Random Projection LSH: We used Bucketed Random Projection LSH to quickly find profiles similar to a user's job interests in our large dataset. This method works well with big data and doesn't use up too much bandwidth or memory. It helped us focus on the top 100 profiles that match a user's job goals, making our job recommendation process more efficient. The outcome is a select group of profiles that are closely related to the job a user wants.

K-Nearest Neighbors (KNN) for Profile Matching: Following the initial refinement with Bucketed Random Projection LSH, we applied a K-Nearest Neighbors (KNN) algorithm for a more granular comparison. This technique evaluates a new user's profile against existing entries in the dataset, considering factors like professional skills, academic qualifications (amount), and work experience (amount). KNN helps identify the profiles most similar to the user's, enabling us to generate tailored insights and recommendations based on comparable career paths.

Gemini for Generating Insights: Gemini, a decoder-only LLM, was utilized for multiple purposes:

- **Extracting Skills from Resumes:** Gemini analyzes resumes uploaded by users to extract and identify their current skill sets, enhancing the personalization of the tool.
- **Generating Actionable Insights:** Based on the skills and job titles, Gemini provides specific recommendations, such as suggesting essential skills that are missing. Additionally, it crafts concise, resume-appropriate descriptions of recommended skills.
- **Recommending LinkedIn Learning Courses:** To address skill gaps, Gemini suggests relevant courses from LinkedIn Learning, further aiding users in their professional development. We supplied Gemini with 750 courses out of our 974 because of max_token_limit limitation.

Evaluation and Results

Our project included a thorough evaluation of the Career Advancement Tool (CAT) and its algorithms. We assessed both the quantitative performance and the qualitative impact of our methods, including KNN and BRP. This evaluation provided important insights and demonstrated the tool's effectiveness.

Neural Network Performance: Our neural network demonstrated strong predictive ability, accurately mapping job titles and company names to relevant skills with **91.75%** accuracy in training and **83.05%** in testing. This high accuracy highlights the network's effectiveness in deciphering complex relationships between job titles and required skills, crucial for customizing user profiles. Key Finding: The neural network's precision in predicting skills highlights how machine learning can significantly improve job marketability by aligning resumes with current industry standards.

Bucketed Random Projection LSH: The application of Bucketed Random Projection LSH (BRP LSH) effectively narrowed down our dataset to profiles closely related to the user's job title, such as "Data Analyst". This filtering step ensured that our analysis was conducted on a relevant and focused subset of data. **Key Finding:** BRP LSH's ability to filter profiles based on job title similarity highlights its value

in pre-selecting the most pertinent profiles for further analysis, streamlining the process of generating personalized recommendations.

KNN Effectiveness: Utilizing the K-Nearest Neighbors (KNN) algorithm on the filtered dataset, we successfully identified the five profiles most akin to the user's, based on degrees, past job experiences, and skills. This facilitated the delivery of deeply personalized career insights.
Key Finding: The KNN algorithm's success in finding similar profiles emphasizes the importance of peer-based insights in personal and professional development, offering users a mirror to the potential pathways their careers could take.

Gemini's Impact: Gemini significantly influenced user engagement by providing actionable insights and recommending LinkedIn Learning courses that matched users' skill gaps. The positive response and improvements seen in user profiles post-engagement with suggested pathways attest to Gemini's effectiveness.

Key Finding: Gemini's impact on user engagement and the tangible improvements in their professional profiles validate the tool's utility in guiding users toward fulfilling their career advancement goals.

Limitation and Reflection

Throughout the development of the Career Advancement Tool (CAT), we encountered several constraints and challenges that shaped our approach and influenced the project's outcomes. This section outlines the key limitations we faced and reflects on their impact.

1. Availability of Skills Data: A major challenge was the absence of specific skill data for employed individuals in our Profiles dataset, which is crucial for our tool's goal to align job seekers with market-demanded skills.

Solution: To overcome this, we utilized a dataset from Kaggle containing job titles, company names, and their required skills. We then developed a neural network to predict necessary skills from these job titles. This method helped us bypass the data shortage, allowing CAT to suggest skills based on the job titles found in user profiles and job postings.

2. Initial Challenges with KNN Implementation: Our initial use of the K-Nearest Neighbors (KNN) algorithm struggled with the complex data structure of high-dimensional vectors for job titles and skills. This complexity reduced its effectiveness, likely due to the "curse of dimensionality."

Solution: We improved our method by introducing a two-step process. First, we used Bucketed Random Projection LSH to filter profiles by job title similarity, which helped manage the high-dimensional data. Then, we applied KNN to align users with profiles, focusing on their educational qualifications, past job experiences, and skills. This approach increased the accuracy of our recommendations, making them more targeted and personalized.

3. Computational Performance Constraints: We faced significant delays in result generation on the Databricks platform due to the extensive running times of our algorithms, which impacted the overall project timeline since each phase depended on the completion of the previous one.

Reflection: Facing challenges highlighted the importance of adaptability and innovation. The absence of detailed skills data inspired us to develop a system that autonomously generates skill recommendations, improving our tool's functionality. Additionally, issues with the initial KNN implementation led to a more efficient two-step matching process, showcasing the value of iterative development. We also recognized the need for better data processing and optimization, providing valuable insights for managing large-scale projects.

4. Gemini's max token limit: The Google Gemini model imposes a maximum token limit, which restricts the length of prompts we can input. This limitation has prevented us from including all 974 courses in our prompt due to the excessive length.

Solution: To comply with Gemini's token limit and ensure our prompt fits within the platform's constraints, we've curated a selection of 750 courses out of the total 974. This approach allows us to work within the token restrictions while still providing a substantial and representative sample of courses for the prompt.

Conclusion

In conclusion, the Career Advancement Tool (CAT) has made significant progress in meeting the dynamic needs of job seekers in a fast-changing job market. Utilizing advanced machine learning algorithms and large language models, CAT effectively bridges the skill gap between users and market demands, offering tailored, data-driven career development solutions.

Key achievements include:

- **Neural Network Development:** This neural network accurately predicts necessary skills from job titles, addressing the issue of unavailable skills data and autonomously enhancing user profiles to meet industry standards.
- **Use of Bucketed Random Projection LSH and KNN:** A two-step process combining these methods has refined the accuracy of our profile matching, allowing CAT to provide more relevant and personalized career insights.
- **Personalized Recommendations through Gemini:** Gemini's role in generating insights and recommending courses has significantly boosted user engagement and empowered users in their career development.

Despite challenges like data limitations and computational issues, our team has shown resilience and creativity, enhancing CAT's functionality and demonstrating the transformative potential of AI and machine learning in career development. These results not only highlight CAT's effectiveness but also pave the way for future improvements and expansions. Our commitment to continuous improvement is driven by the belief in technology's role in advancing professional development and helping individuals achieve their career goals.

Appendix

Images, Graphs, Plots:

BRP LSH output:

Table

New result table: ON

Search

	position	job_title	education_count	companies_count	job_title_embeddings	skills_vector	skills
1	Data Analyst	data analyst	4	1	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","emr","git
2	Data Analyst	data analyst	2	4	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er
3	Data Analyst	data analyst	1	7	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er
4	Data Analyst	data analyst	3	3	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data modeling",
5	Data Analyst	data analyst	1	7	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data modeling","data sc
6	Data Analyst	data analyst	1	5	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","emr","git
7	Data Analyst	data analyst	0	5	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","emr","git
8	Data Analyst	data analyst	3	2	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","emr","git
9	Data Analyst	data analyst	2	0	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data modeling",
10	Data Analyst	data analyst	1	8	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er
11	Data Analyst	data analyst	2	1	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er
12	Data Analyst	data analyst	2	5	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er
13	Data Analyst	data analyst	3	4	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","git","lead
14	Data Analyst	data analyst	2	0	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","er

KNN output:

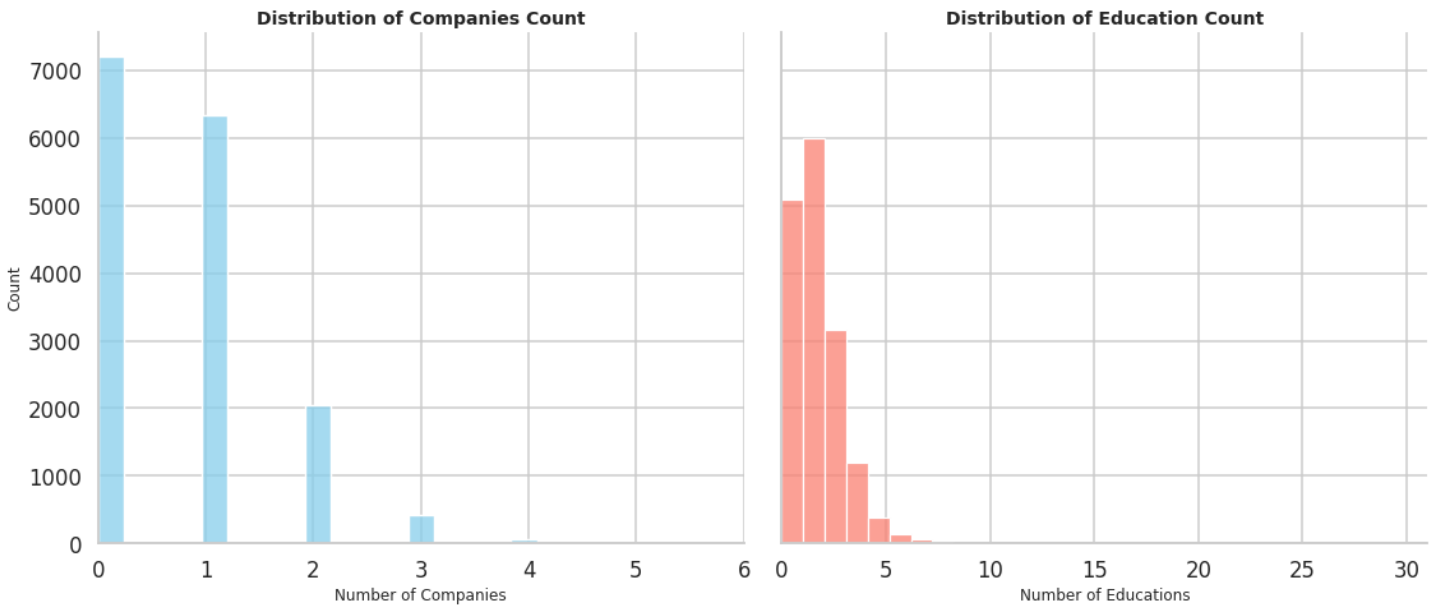
Table

New result table: ON

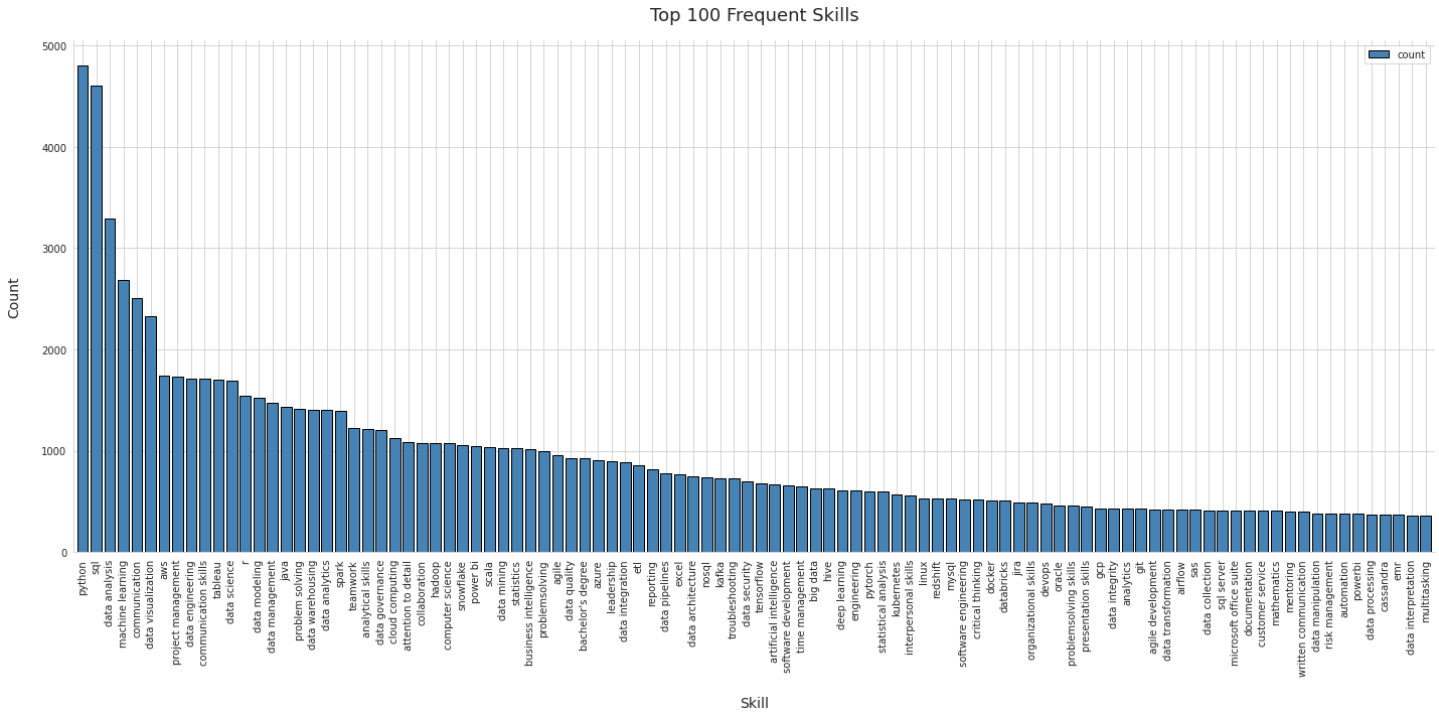
Search

	position	job_title	education_count	companies_count	job_title_embeddings	skills_vector	skills
1	Data Analyst	data analyst	2	4	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","emr","git",
2	Data Analyst	data analyst	2	6	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","emr","git",
3	Data Analyst	data analyst	2	5	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data modeling","data scier
4	Data Analyst	data analyst	2	5	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","emr","git",
5	Data Analyst	data analyst	2	4	[-0.19284888,-0.093611...	[0,0,0,0,0,0,0,...	["agile","automation","cassandra","data integrity","data management","data science","git","leader

Feature distributions:



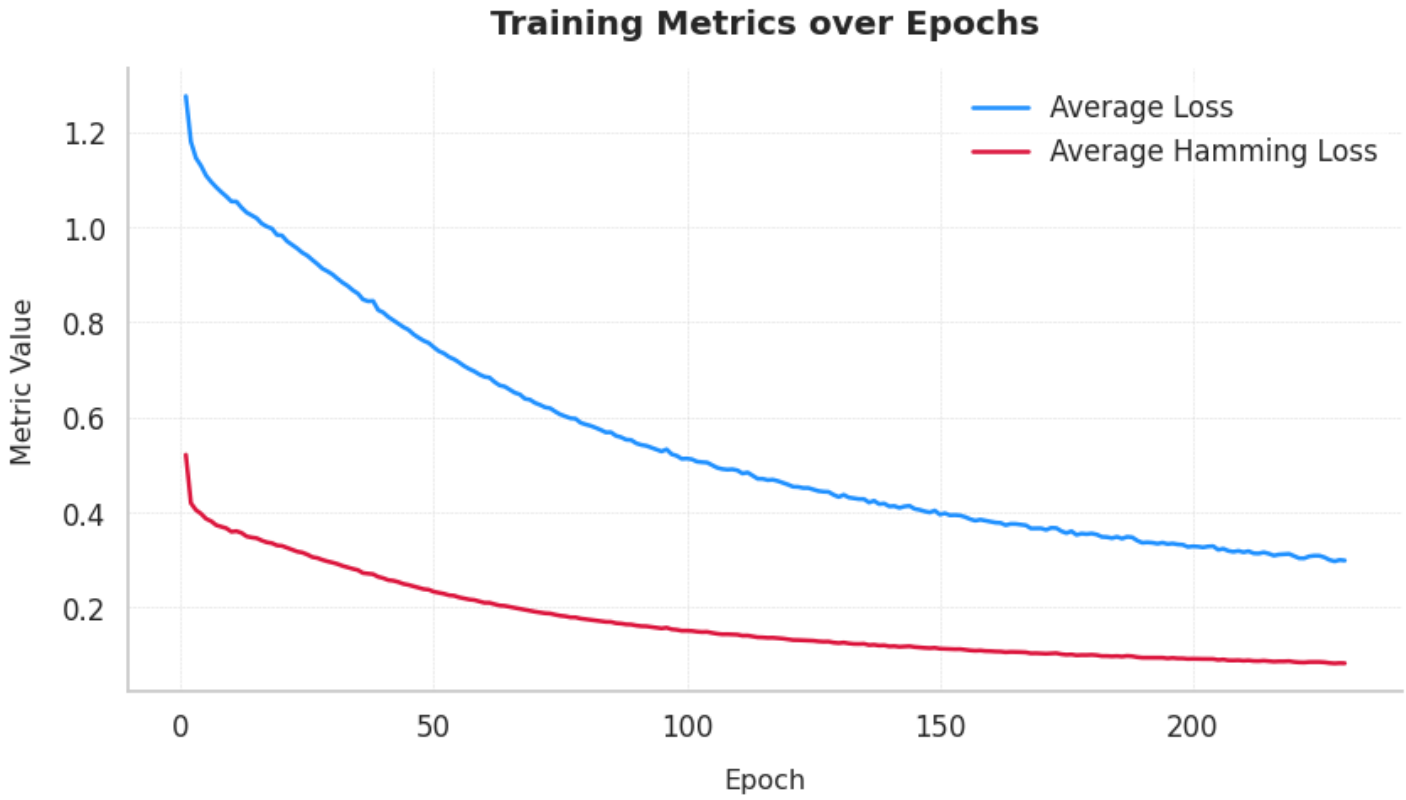
Top frequent skills:



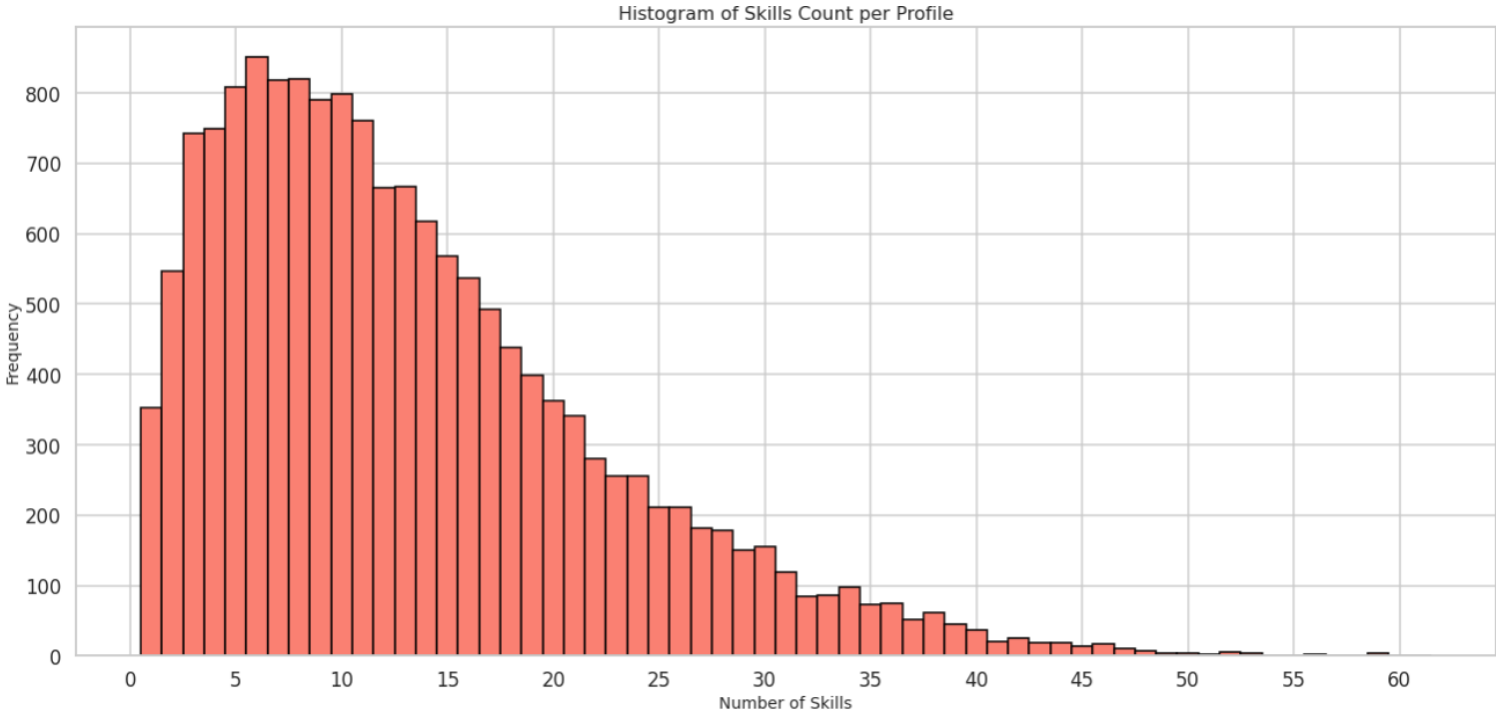
Job postings and skills:

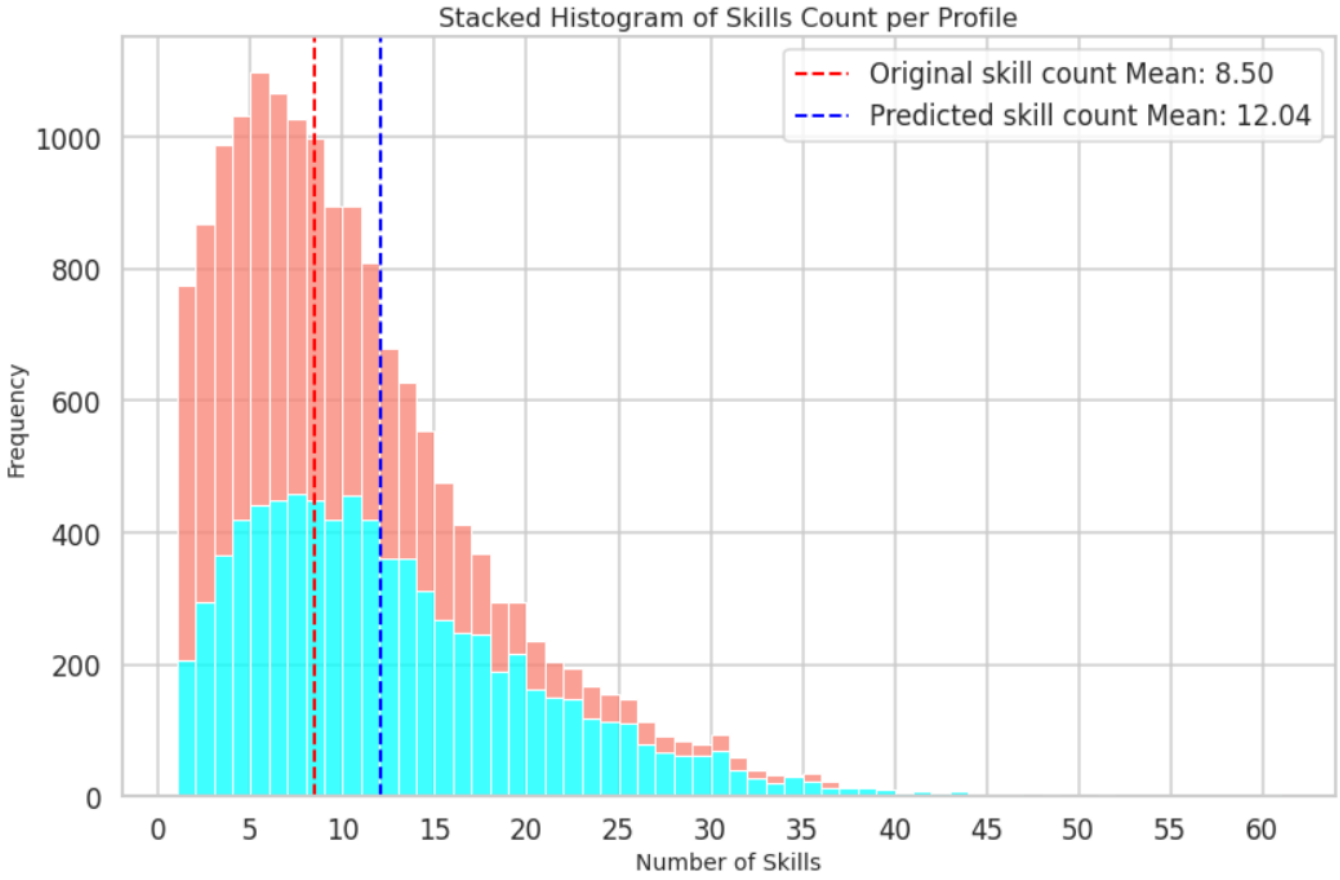
position	job_title	education_count	companies_count	job_title_embeddings	skills_vector
CPA| Analyst	cpa analyst	0	4	[-0.26191905, -0....	[0, 0, 0, 0, 0, 0...
Director Data & A...	director data an...	2	6	[-0.30737558, -0....	[0, 0, 0, 0, 0, 0...
Highly experience...	highly experience...	3	9	[-0.33366272, 0.0...	[0, 0, 0, 0, 0, 0...
SEO Analyst at Le...	seo analyst at le...	1	2	[-0.3200069, -0.1...	[0, 0, 0, 0, 0, 0...
Senior Data Analy...	senior data analy...	3	3	[-0.4128574, -0.1...	[0, 0, 0, 0, 0, 0...
Product Managemen...	product managemen...	4	2	[-0.30798188, -0....	[0, 0, 0, 0, 1, 0...
Creator Success |...	creator success ...	2	12	[-0.33328816, -0....	[0, 0, 0, 0, 0, 1...
Experienced Prope...	experienced prope...	2	4	[-0.25802732, 0.0...	[0, 0, 0, 0, 0, 0...
Senior Tax Analys...	senior tax analys...	2	0	[-0.3276589, -0.1...	[0, 0, 0, 0, 1, 0...
Strategic and Dat...	strategic and dat...	3	17	[-0.3889637, 7.04...	[0, 0, 0, 0, 0, 0...
Senior Financial ...	senior financial ...	2	4	[-0.401947, 0.019...	[0, 0, 0, 0, 0, 0...
Business Systems ...	business systems ...	1	1	[-0.29397982, -0....	[0, 0, 0, 0, 0, 0...
Operations Resear...	operations resear...	2	7	[-0.3203108, -0.1...	[0, 0, 0, 0, 0, 0...
Give me the data ...	give me the data ...	4	4	[-0.075327404, -0...	[0, 0, 0, 0, 0, 0...
Data and AI | Arc...	data and ai arch...	3	12	[-0.36926755, -0....	[0, 0, 0, 0, 0, 0...
Senior Equity Res...	senior equity res...	1	5	[-0.34441507, -0....	[0, 0, 0, 0, 1, 1...
Marketing Analyst...	marketing analyst...	2	4	[-0.24074626, 0.0...	[0, 0, 1, 0, 1, 0...
Integrated Media ...	integrated media ...	2	8	[-0.28109893, -0....	[0, 0, 1, 0, 1, 0...

Neural Network training:



Distribution of skill count per profile (how many skills each person have)



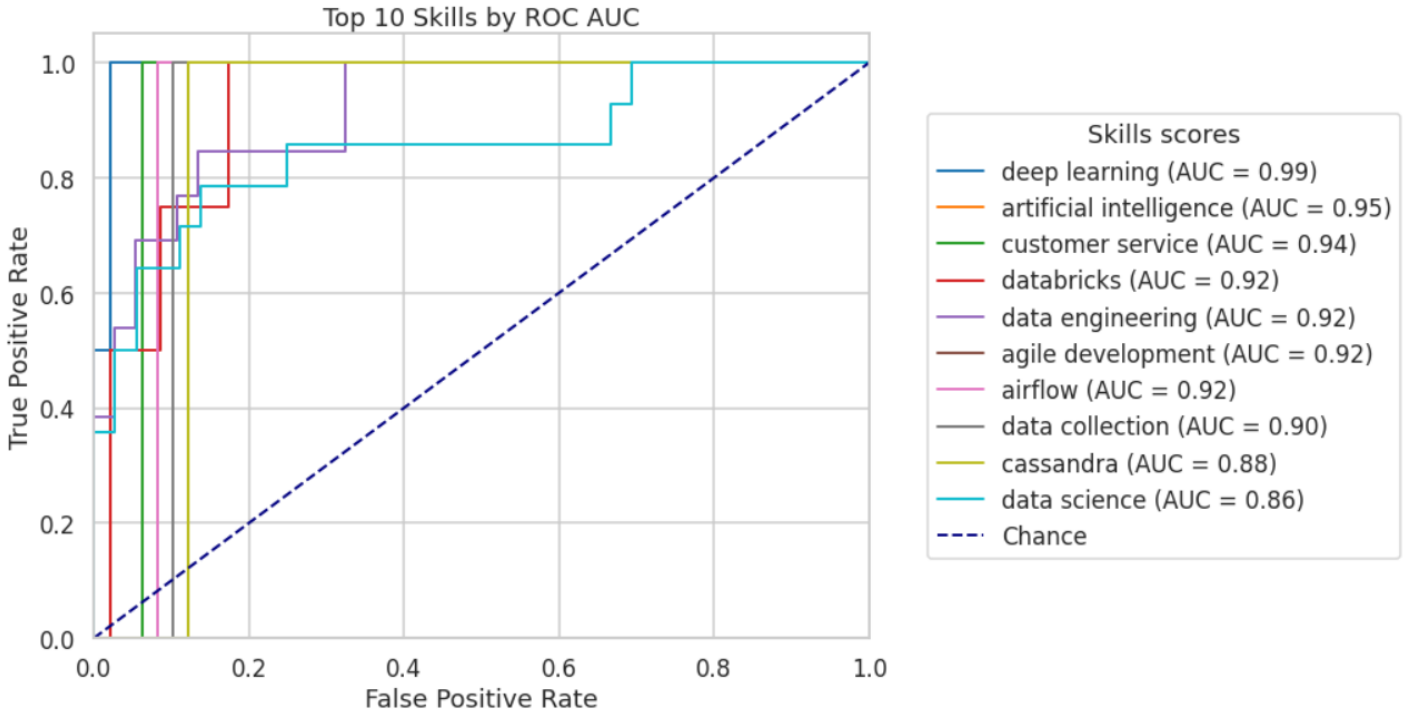


Gemini insights output (A snippet output for a designated user):

I found the following courses that are the most relevant to the skills you listed:

- **Leadership:** Leading with Vision
  - I am able to inspire others to achieve a common goal and create a positive and productive work environment.
- **Mathematics:** Statistics Foundations 1: The Basics
  - I am able to analyze and interpret data, and use statistical methods to solve problems.
- **Presentation skills:** Presenting Data Effectively to Inform and Inspire
  - I am able to create and deliver presentations that are clear, concise, and engaging.
- **Data integration:** Introduction to Data Warehouses
  - I am able to understand the principles of data warehousing and how to integrate data from different sources.
- **Problemsolving:** Problem-Solving Techniques
  - I am able to identify and solve problems using a variety of techniques, including critical thinking and analytical reasoning.
- **SQL server:** SQL Server Fundamentals: Master Basic Query Techniques
  - I am able to write and execute SQL queries to retrieve and manipulate data from a SQL Server database.
- **R:** R for Data Science: Analysis and Visualization
  - I am able to use R to analyze and visualize data, and create statistical models.
- **"bachelor's degree":** [Placeholder]
  - I have earned a bachelor's degree in a relevant field.
- **Data warehousing:** Introduction to Data Warehouses
  - I am able to understand the principles of data warehousing and how to design and implement a data warehouse.
- **Data collection:** Data Collection Essentials
  - I am able to collect data from a variety of sources, including surveys, interviews, and observations.
- **Communication:** Communicating with Confidence
  - I am able to communicate effectively with people from all levels of an organization, both verbally and in writing.
- **Communication skills:** Communication Skills for Modern Management
  - I have strong communication skills, including the ability to listen, speak, and write effectively.
- **Data interpretation:** Data Interpretation for Business Professionals
  - I am able to interpret data and identify trends and patterns.

Top 10 Skills with the Highest AUC Scores:



Top 10 Skills with the Lowest AUC Scores: