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 project, titled **Career Advancement Tool (CAT)**, is designed as a cutting-edge feature within LinkedIn, aimed at augmenting the marketability of job seekers. This innovation emerges from the recognition of a critical gap between the existing skill sets presented in users' profiles and the evolving demands of the job market. CAT seeks to address this disparity by leveraging Machine Learning and LLM to both analyze and suggest enhancements to resumes, ensuring they align with current industry standards. Additionally, it aims to guide users in acquiring new skills or improving existing ones to meet these standards.

The necessity for CAT stems from a dynamic job market where the requirements for positions are constantly changing. Traditional methods of job seeking and skill presentation are becoming less effective, leaving many qualified candidates unnoticed or unconsidered. CAT responds to this challenge by offering a personalized, data-driven approach to professional development. It not only helps users to refine their resumes but also identifies skill gaps and recommends targeted learning pathways through LinkedIn Learning courses. This dual approach ensures that users' profiles not only meet but exceed the current industry standards, significantly increasing their chances of job match success and fostering ongoing 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

To enhance the depth and relevance of the Career Advancement Tool (CAT), our team has augmented the foundational LinkedIn dataset with up-to-date information on job market demands, including required skills and qualifications. Additionally, we have sourced data regarding courses and certifications from LinkedIn Learning, an esteemed online platform offering a wide array of video courses taught by industry experts across various domains such as software development, creative design, and business management.

This augmentation was accomplished through two primary methods: web scraping for current courses from LinkedIn Learning and incorporating a publicly available, pre-scraped job skills dataset from Kaggle. The Kaggle dataset, specifically scraped and uploaded in 2024, includes detailed information about skills and qualifications sought in the job market. Our web scraping effort focused on extracting the latest courses and certifications from LinkedIn Learning, ensuring that the CAT can recommend the most relevant and recent learning opportunities to its users.

Given the project's constraints and the large spectrum of job sectors, our data collection efforts were strategically concentrated on the data science sector. This focus allows us to provide targeted and highly relevant recommendations for individuals seeking to advance in or transition to careers within this rapidly growing field. By integrating this carefully selected additional data, CAT is equipped to offer insights that reflect the latest trends and requirements of the job market, thereby enhancing its effectiveness in guiding users towards successful career development.

Integration of Additional Data

The integration of additional data into the Career Advancement Tool (CAT) significantly enhances the utility and accuracy of its recommendations. This process involves the merging of dynamically sourced data with the foundational LinkedIn profile dataset.

A notable step in refining our dataset was the analysis of the vast array of skills identified from the Kaggle dataset. With over 10,000 unique skills cataloged, it became essential to simplify this list to focus on the most impactful and frequently demanded skills within the job market. Consequently, we narrowed our focus to the top 100 skills (this was done by filtering the jobs skills dataset), as depicted in Figures 1 and 2. This curation ensures that CAT's

recommendations remain relevant and targeted to the core competencies valued across industries.

To further enhance our model's predictive capabilities, we utilized the dataset of skills from Kaggle to train a **neural network**. This network is designed to predict a multi-hot vector, which represents a set of skills derived from an input job title and company (the company which posted the job in LinkedIn) strings.
The performance of this neural network was carefully evaluated, achieving a Hamming loss of 9.31% (Which means accuracy of hamming distance is 90.69%) on the training set and 17.71% (Which means accuracy of hamming distance is 82.29%) on the test set, with the data being split **80%/20%** between training and testing, respectively. This evaluation indicates a pretty robust model capable of accurately predicting skill sets from job titles.

Leveraging this trained neural network, we were able to augment the LinkedIn Profiles dataset by predicting and assigning relevant skills to each profile based on the job title of the individual's current position. This smart integration method not only enriches the dataset but also enables CAT to offer more nuanced and precise guidance to users, customizing recommendations to their specific professional experiences and aspirations.

Items and Enrichment Size

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

- For the LinkedIn Profiles dataset, an item represents a user's profile, which includes professional experiences, skills, educational background, and current position.
- For the LinkedIn Learning courses data, gathered through web scraping, an item refers to a course offered on the platform. Each course item includes the course name, its rating, and the number of individuals who have participated in the course.
- For the job postings data, an item constitutes an individual job listing. The details for each job listing item cover the job title and the set of skills required for that particular job.

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

Utilizing Bucketed Random Projection LSH, designed for high-dimensional data, we efficiently identified approximate nearest neighbors within our large Profiles dataset. This algorithm excels in big data scenarios thanks to Spark's distributed computing, enabling parallel processing without requiring data broadcast, saving bandwidth and memory. It proved effective in narrowing down to the top 100 profiles most closely aligned with a new user’s job preferences, optimizing our matchmaking process between users and job opportunities.
The result of this part is a subset of the Profiles dataset, that holds profiles that holds a position (an embedding of the position) that is closes to most to the new user’s desired job.

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's comprehensive evaluation aimed to validate the effectiveness of the Career Advancement Tool (CAT) and its underlying algorithms. Through this process, we conducted quantitative assessments of performance while also evaluating the qualitative outcomes of our algorithms, such as the KNN and BRP, leading to significant insights and results.

Neural Network Performance

The neural network showcased strong predictive performance in mapping job titles and publishers (company name) to relevant skills, achieving an accuracy of **91.75%** on the training set and **83.05%** on the **test set**, based on the Hamming loss metric. This high level of accuracy demonstrates the neural network's capability to understand and process the complex relationships between job titles and publishers (company name) to their associated skills, a cornerstone for tailoring user profiles.

Key Finding: The neural network's precision in skill prediction underscores the potential of machine learning to revolutionize job marketability by enhancing resume relevance to 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

One of the primary challenges was the lack of available data on the skills of individuals who are currently employed (items in Profiles dataset), a core aspect essential to fulfilling the project's main objective. This limitation presented a significant hurdle, as our tool aims to match job seekers with skills that are in demand within the job market.

Solution: To address this gap, we used a scrapped data from Kaggle of job titles, publishers (company name) and their required skills and then we developed a neural network capable of predicting necessary skills based on job titles. This innovative approach allowed us to circumvent the lack of direct skills data, enabling CAT to generate skill recommendations by analyzing the job titles present in user profiles and available job postings.

2. Initial Challenges with KNN Implementation

Our initial implementation of the K-Nearest Neighbors (KNN) algorithm faced difficulties due to the inclusion of high-dimensional vectors representing job title embeddings and skills sets in the new user's and profiles' representing vectors. This complexity hindered the algorithm's effectiveness in matching users with similar profiles based on career-related factors, probably due to the "curse of dimensionality".

Solution: We refined our approach by adopting a two-step process. First, we applied Bucketed Random Projection LSH to filter profiles based on job title similarity, addressing the issue of high dimensionality. Following this, we utilized KNN to match users with profiles based on the number of degrees, past jobs, and their shared skills. This sequential method improved the precision of our matching process, enabling more relevant and personalized recommendations.

3. Computational Performance Constraints

Another significant challenge was the extensive running times experienced on the Databricks platform, which delayed the generation of results critical to the subsequent stages of our project. Given the linear progression of our algorithms, delays in one phase affected the timely execution and evaluation of the next.

Reflection: These challenges highlighted the importance of adaptability and innovation in overcoming project constraints. The limitation regarding skills data prompted us to develop a solution that not only addressed the gap but also enhanced the project's capability to generate skill recommendations autonomously. The initial setback with the KNN algorithm led to a more refined and effective two-step matching process, demonstrating the value of iterative development and problem-solving.

The computational performance constraints underscored the need for efficient data processing and algorithm optimization, especially when dealing with large datasets and complex analyses. This experience has provided valuable insights into managing project timelines and resource allocation in data-intensive projects.

Overall, while these limitations posed significant challenges, they also prompted us to explore creative solutions and refine our methodologies, ultimately contributing to the development of a more robust and effective career advancement tool. Reflecting on these experiences, we recognize the importance of flexibility, problem-solving, and continuous improvement in the face of project constraints, lessons that will inform our future endeavors in the field of data science and machine learning.

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 development of the Career Advancement Tool (CAT) represents a significant stride towards addressing the dynamic needs of job seekers in today's rapidly evolving job market. By leveraging advanced machine learning algorithms and large language models, CAT has successfully bridged the gap between the existing skills of users and the demands of the marketplace, offering a personalized, data-driven approach to career development.

Key achievements of the project include:

- **Development of a Neural Network:** This neural network predicts necessary skills from job titles with high accuracy, overcoming the challenge of unavailable skills data. This core component ensures that CAT can autonomously generate skill recommendations, enhancing user profiles to align with industry standards.
- **Effective Use of Bucketed Random Projection LSH and KNN:** The implementation of a two-step process combining BRP LSH and KNN for profile matching significantly improved the precision of our recommendations. This methodological refinement has enabled CAT to offer more relevant and personalized career insights based on career-related factors such as the number of degrees, past job experiences, and skills.
- **Personalized Recommendations through Gemini:** The use of Gemini for generating actionable insights and recommending LinkedIn Learning courses based on identified

skill gaps has markedly increased user engagement. This not only aids users in their professional development but also empowers them to take proactive steps towards achieving their career goals.

Despite facing challenges such as data limitations and computational constraints, the project team demonstrated resilience and ingenuity by developing creative solutions that not only addressed these issues but also enhanced the overall functionality of CAT.

The outcomes of this project underscore the potential of AI and machine learning technologies to revolutionize career development services. CAT's success in providing personalized, actionable guidance to job seekers reflects its capability to become an essential tool in navigating the complexities of the modern job market.

As we look to the future, the insights gained from this project will serve as a foundation for further enhancements and expansions of CAT. The team is committed to continuous improvement, driven by the belief in the power of technology to transform professional development and enable individuals to realize their full potential in their careers.

Appendix

Images, Graphs, Plots:

BRP LSH output:

Table | New result table: ON | Search

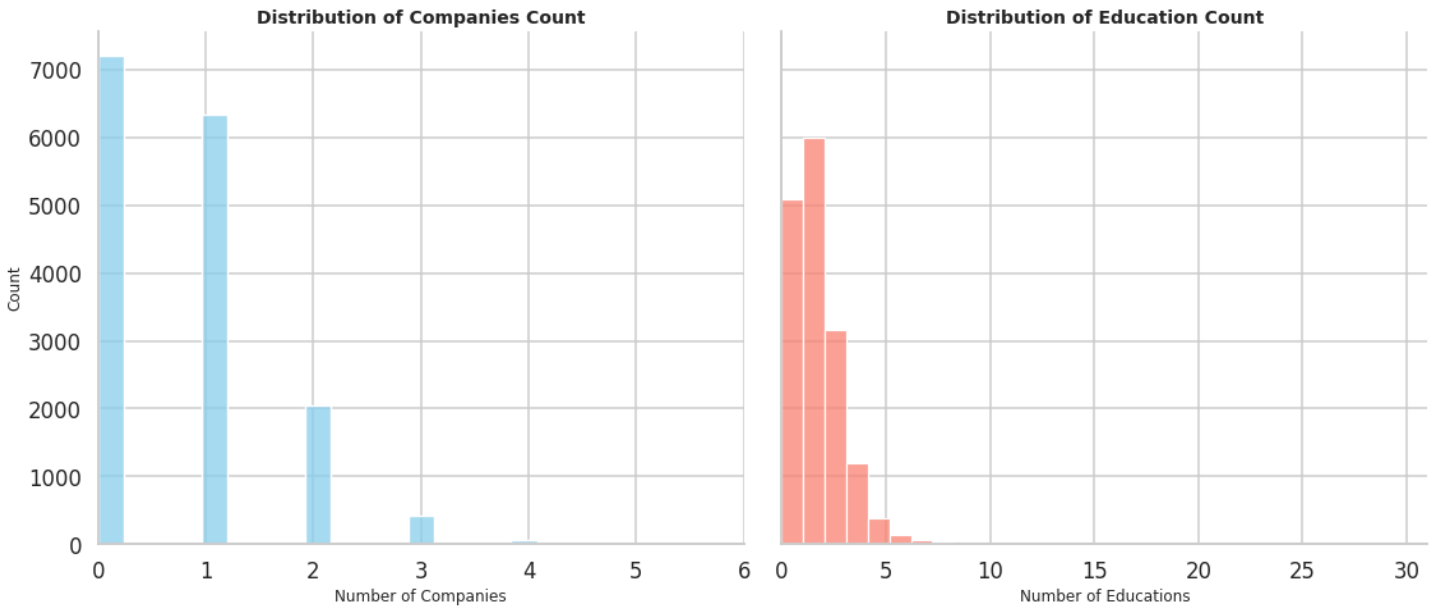
	position	job_title	education_count	companies_count	job_title_embeddings	skills_vector	skills
1	Data Analyst	data analyst	4	1	[-0.19284888,-0.093611...	[1,0,0,0,0,0,0,...	["automation","cassandra","data integrity","data management","data science","emr","git
2	Data Analyst	data analyst	2	4	[-0.19284888,-0.093611...	[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:

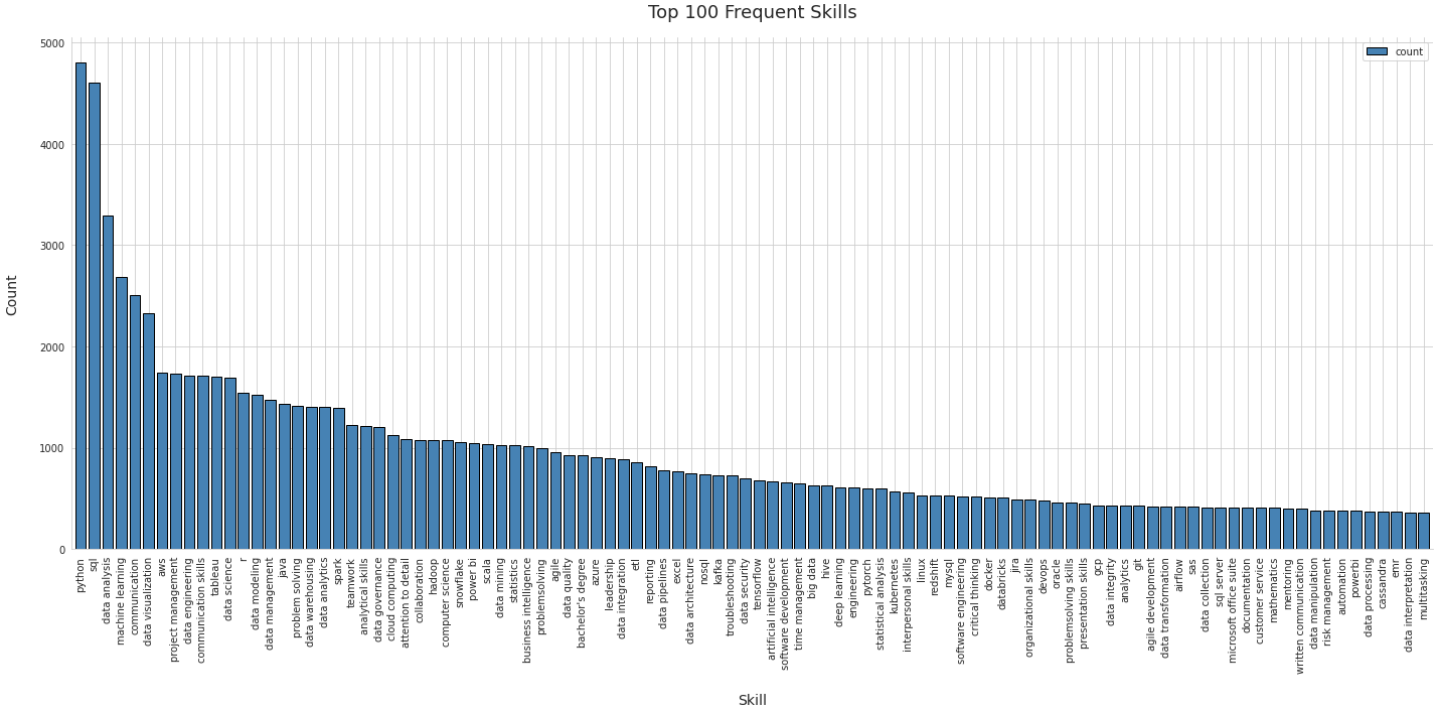
Table | New result table: ON | Search

	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","l
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","l
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","l
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","leaders

Feature distributions:



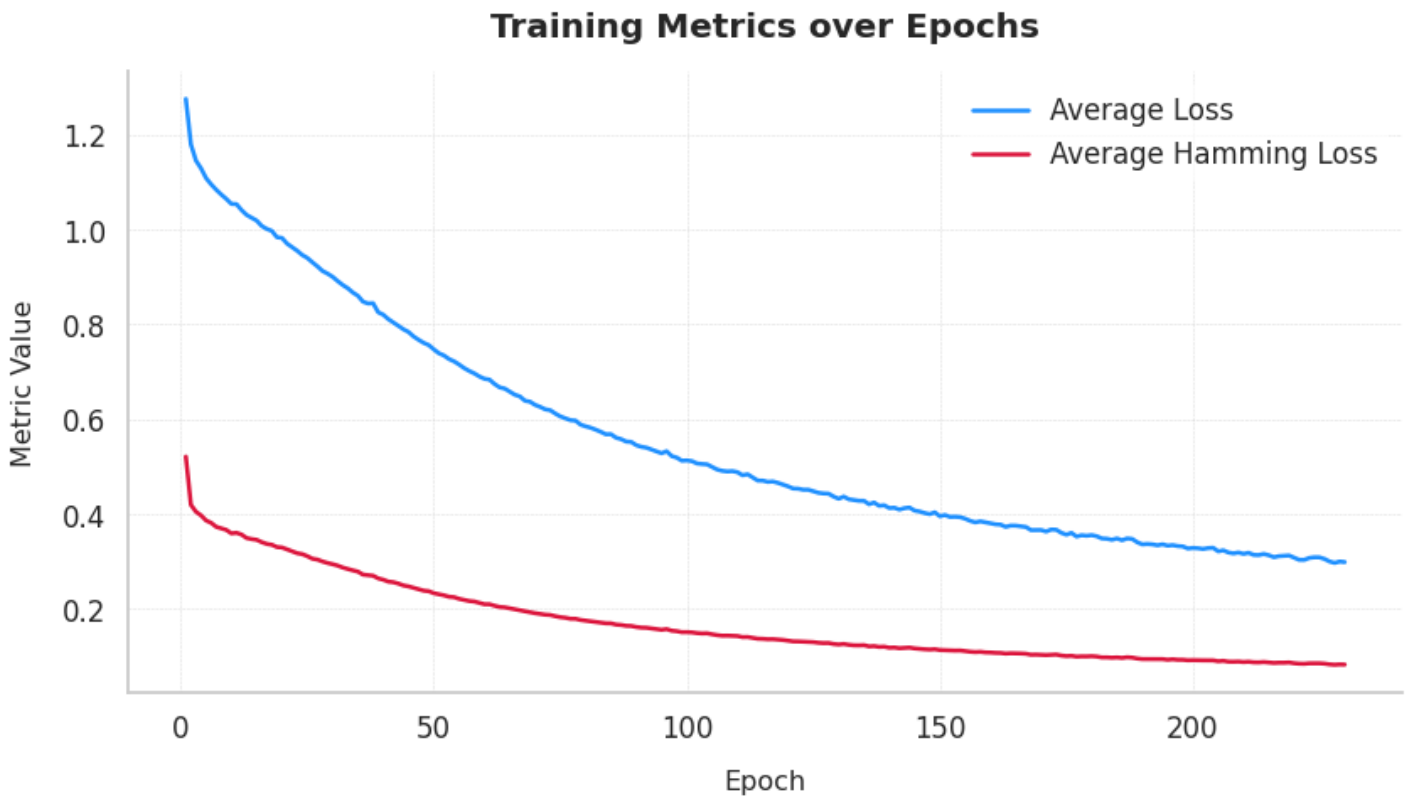
Top frequent skills:



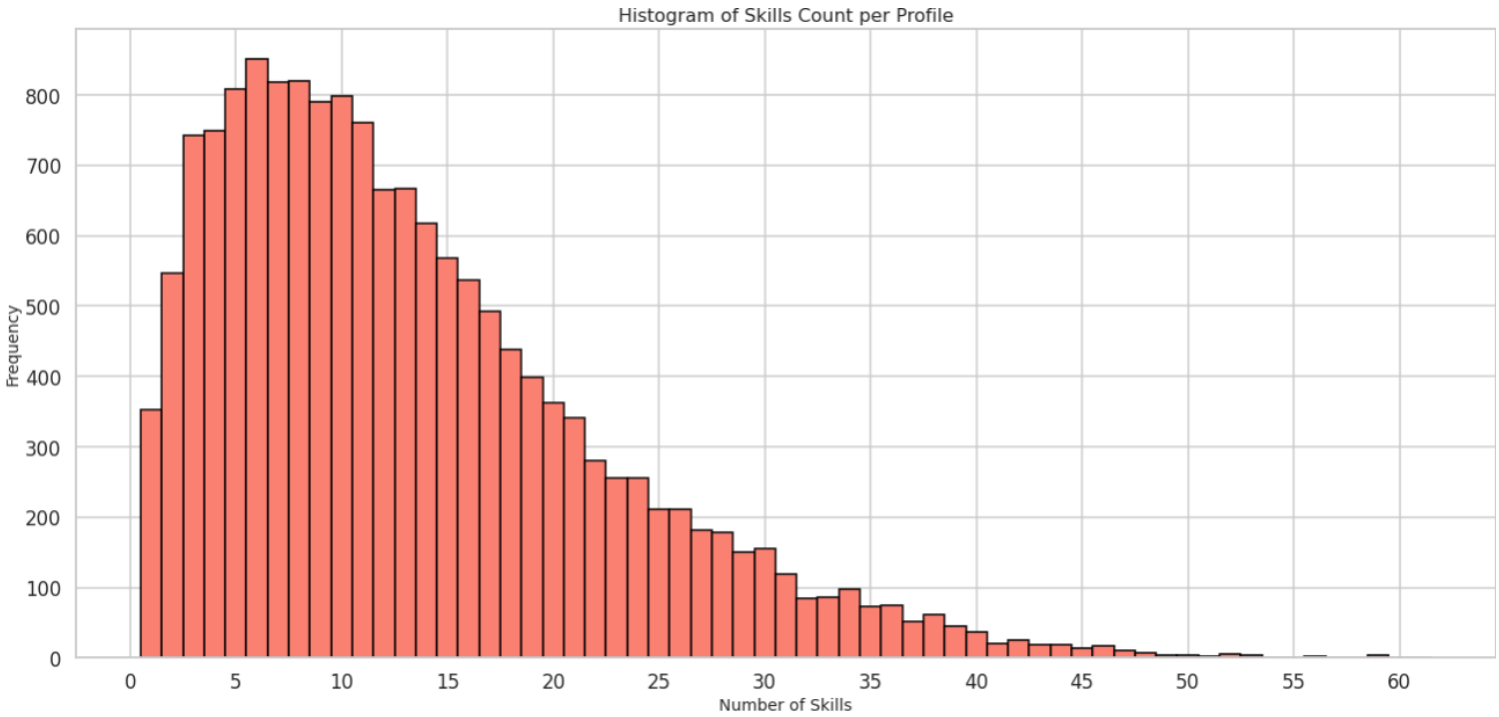
Job postings and skills:

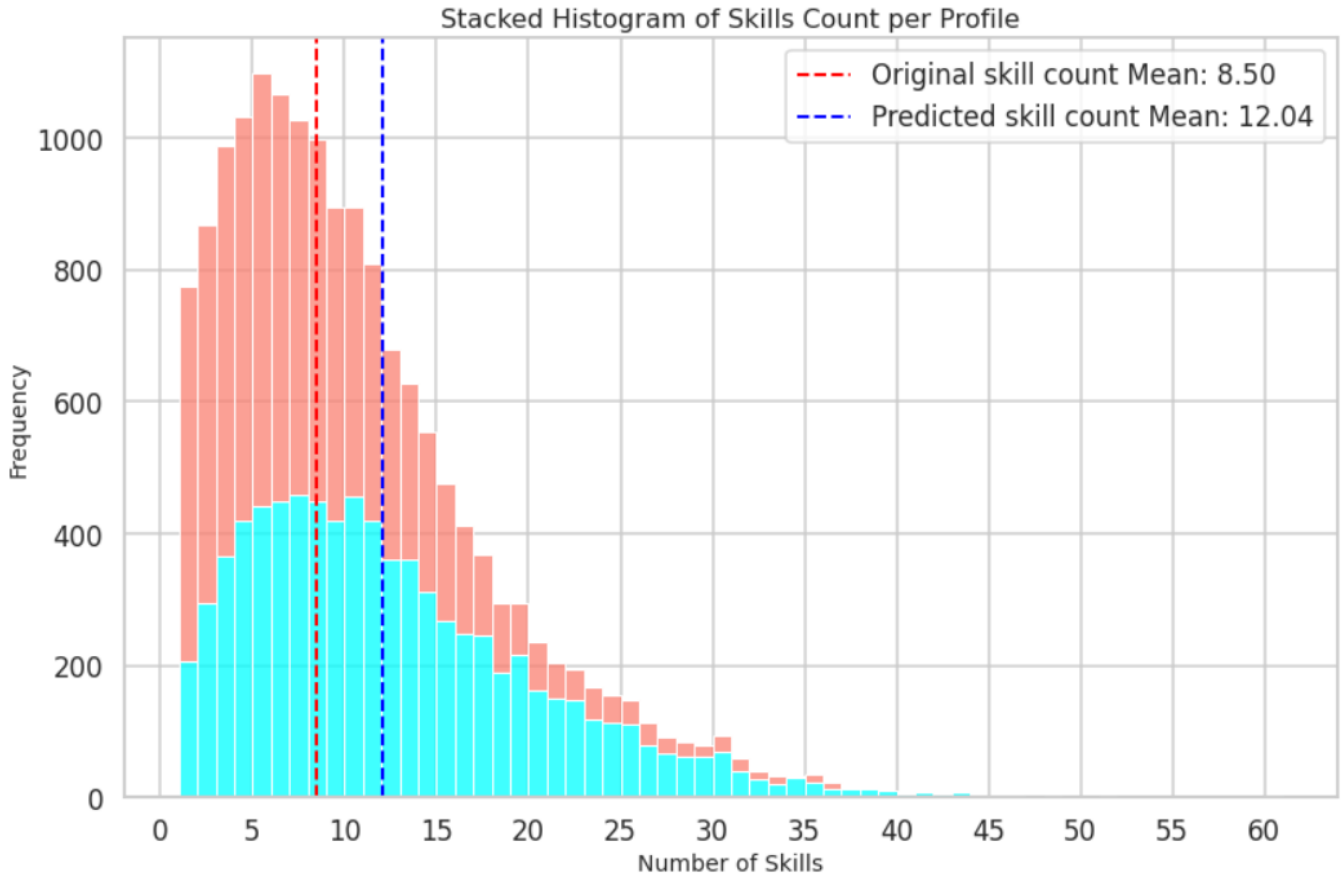
position	job_title	education_count	companies_count	job_title_embeddings	skills_vector
CPA| Analyst	cpa analyst	0	4	[-0.26191905, -0....	[0, 0, 0, 0, 0, 0...
Director Data & A...	director data an...	2	6	[-0.30737558, -0....	[0, 0, 0, 0, 0, 0...
Highly experience...	highly experience...	3	9	[-0.33366272, 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)



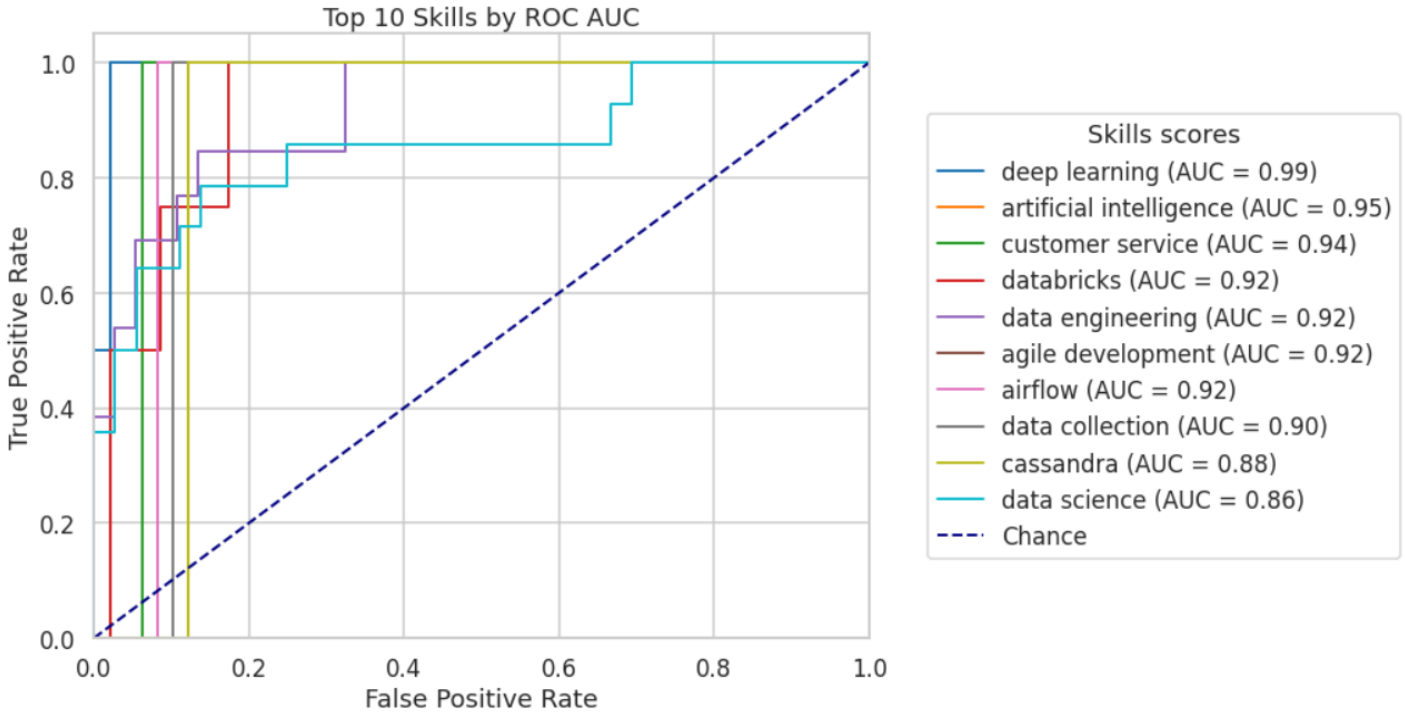


Gemini insights output (A snippet output for a designated user):

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: