Project Report

Course: Data Collection lab - 094290

Participants: Luay Marzok, Rany Khirbawy, Weaam Mulla

Deadline: 09/04/2024

Table Of Contents

Project Introduction:

In today's rapidly evolving job market, finding the right career path can be daunting. That's why we've developed a personalized career pathing solution powered by advanced AI algorithms and comprehensive data analysis.

Inspired by insights from platforms like LinkedIn and Glassdoor, our project aims to answer a crucial question: Can we predict an individual's optimal career path using AI-driven analysis of their personal attributes and current job market data?

This project addresses the critical need for tailored career guidance in an increasingly complex professional landscape, offering individuals clarity and confidence in their career decisions.

Data Collection and Integration:

Our analysis relied mainly on the profiles dataset, as our AI solution is specifically tailored for LinkedIn users.

A web scraping script was developed using Python's Selenium and BeautifulSoup libraries. The script navigates through Glassdoor's website, specifically targeting job listings.

Selenium was utilized to automate the web browsing process, while BeautifulSoup facilitated the extraction of relevant information from the HTML content.

Once the data is collected, it is organized into a Pandas and saved as a CSV file.

The scraped data consists of job listings extracted from Glassdoor, focusing on the United States job market.

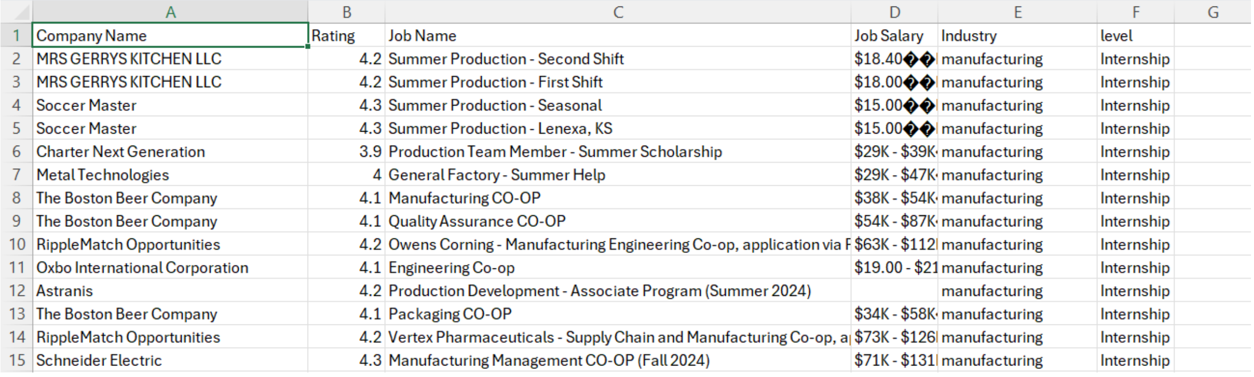
Each job listing contains six features: Company Name, Rating, Job Name, Job Salary, Industry and the Seniority Level Required. The Seniority Level Required serves as a key parameter for categorizing job listings into different career stages, including Internship, Entry Level, Mid Senior, Director, and Executive positions.

Definition of Item:

The total number of job listings (items) scraped, is approximately 4867 listings with six columns, including 'Company Name', 'Rating', 'Job Name', 'Job Salary', 'Industry', and 'Seniority Level Required'.

a(group) ????

Image of scraped data :



	A	B	C	D	E	F	G
1	Company Name	Rating	Job Name	Job Salary	Industry	level	
2	MRS GERRYS KITCHEN LLC	4.2	Summer Production - Second Shift	$18.40��	manufacturing	Internship	
3	MRS GERRYS KITCHEN LLC	4.2	Summer Production - First Shift	$18.00��	manufacturing	Internship	
4	Soccer Master	4.3	Summer Production - Seasonal	$15.00��	manufacturing	Internship	
5	Soccer Master	4.3	Summer Production - Lenexa, KS	$15.00��	manufacturing	Internship	
6	Charter Next Generation	3.9	Production Team Member - Summer Scholarship	$29K - $39K	manufacturing	Internship	
7	Metal Technologies	4	General Factory - Summer Help	$29K - $47K	manufacturing	Internship	
8	The Boston Beer Company	4.1	Manufacturing CO-OP	$38K - $54K	manufacturing	Internship	
9	The Boston Beer Company	4.1	Quality Assurance CO-OP	$54K - $87K	manufacturing	Internship	
10	RippleMatch Opportunities	4.2	Owens Corning - Manufacturing Engineering Co-op, application via F	$63K - $112	manufacturing	Internship	
11	Oxbo International Corporation	4.1	Engineering Co-op	$19.00 - $21	manufacturing	Internship	
12	Astranis	4.2	Production Development - Associate Program (Summer 2024)		manufacturing	Internship	
13	The Boston Beer Company	4.1	Packaging CO-OP	$34K - $58K	manufacturing	Internship	
14	RippleMatch Opportunities	4.2	Vertex Pharmaceuticals - Supply Chain and Manufacturing Co-op, a	$73K - $126	manufacturing	Internship	
15	Schneider Electric	4.3	Manufacturing Management CO-OP (Fall 2024)	$71K - $131	manufacturing	Internship	

Data Analysis:

Analysis Techniques:

Keyword Identification: This technique was used to parse through the text in the 'position' column to identify specific keywords or phrases that are indicative of seniority levels or industries.

Text Analysis: Text analysis techniques, such as sentiment analysis or topic modeling, was applied to columns like 'about' and 'experience' to extract valuable insights.

Statistical Analysis: Statistical techniques, such as calculating distributions, help understand the distribution of education years and experiences among individuals. This analysis provides insights into common educational backgrounds and levels of professional experience within the dataset.

Word Frequency Analysis: Analysing the frequency of words or phrases within the 'description' subcolumn of the 'experience' column can reveal recurring themes or topics across individuals' experiences.

Categorical Analysis: This technique involves categorizing and analyzing the causes listed in the 'cause' column from the 'volunteer_experience.' By aggregating and counting the occurrences of different causes, you can identify prevalent interests or values among individuals.

Feature Selection:

Position Column: Keywords and patterns identified in the 'position' column serve as features for labeling train data. By selecting relevant keywords indicative of seniority levels or industries, you can create a categorical feature that contributes to predicting individuals' profiles accurately.

Education and Experience Columns: The number of years of education and experiences are selected as direct indicators of expertise and seniority level. These numerical features provide quantitative measures of individuals' educational backgrounds and professional experiences, which are crucial for our predictive model.

Field Sub-column within Education: Extracting information about subjects studied during academic journeys serves as features for understanding individuals' areas of expertise.

Description Sub-column within Experience: Common words or themes identified in the 'description' sub-column are selected as features to capture the essence of individuals' experiences.

Cause Column from Volunteer Experience: Causes individuals support through volunteering are selected as features to understand their interests thus predicting the field they belong to.

Certifications Column: Types of certifications held by individuals are selected as features to understand their fields of expertise and proficiency levels.

AI Methodologies:

Regarding the AI methodologies, starting with the deployment of web scraping techniques to aggregate job listings data from Glassdoor, which set the groundwork for our predictive models. The approach was characterized by an extensive data preprocessing phase, where techniques from natural language processing (NLP) were utilized to refine raw profile texts into structured, machine-learning-ready formats and that was done using BERT pre-trained model. This process involved feature engineering to identify and extract pertinent attributes from profiles, such as 'position', 'education', and 'experience', ensuring the data was primed for analysis. Our strategy embraced supervised learning model, logistic regression within an NLP pipeline which was the chosen predictive model after trying to employ different models, leveraging the power of BERT embeddings to deeply understand the context of profile texts for classifying the industry field and seniority level of individuals accurately.

To complement the predictive modeling, our project applied unsupervised learning algorithms for the generation of optimal career paths, marking a significant leap in personalizing career guidance. By converting job names and individual positions into vectors via Word2Vec and applying cosine similarity, our methodology adeptly matched LinkedIn profiles with suitable job opportunities, considering factors such as job ratings and salaries. This nuanced approach not only demonstrated the seamless integration of different AI techniques to address complex career prediction challenges but also underscored the project's commitment to iterative model improvement and evaluation.

Despite the inherent challenges in dealing with incomplete information and outlier profiles, our project's AI methodologies exhibited a robust framework for developing dynamic, AI-driven solutions for career path prediction, showcasing the potential for future advancements in personalized career development tools.

Evaluation and Results:

Our final algorithm has yielded promising results in line with our expectations. It generates a list comprising an optimal career path tailored for each individual based on various criteria such as their current position, seniority level, field of expertise, and the contents of their professional profile. Moreover, the algorithm operates dynamically, adjusting the career trajectory according to the individual's seniority level. For instance, when presented with a person at the "internship" seniority level, it constructs a career path extending to the highest echelon, namely "Director".

We have observed that a significant proportion of these career paths are highly suitable and remarkably accurate. We are exceptionally pleased with the outcomes, underscoring the efficacy of our efforts in algorithm development and refinement.

As with any model, particularly within the realm of artificial intelligence, our algorithm exhibits limitations in accuracy. For instance, instances where individuals provide incomplete information may result in the prediction of career paths that are less than optimal or deviate from ideal roles. This outcome was anticipated given the reliance of our model on comprehensive data inputs for accurate predictions.

Furthermore, individuals occupying roles that are uncommon or considered outliers pose challenges to the model's predictive capabilities. Such cases may lead to deviations from expected outcomes, as these scenarios fall outside the scope of typical data patterns used for training the algorithm.

These examples underscore the inherent limitations of our model and highlight the need for continued refinement and adaptation to address diverse and nuanced scenarios within the realm of career prediction.

Limitations and Reflection:

Throughout our project, we encountered several constraints and challenges that shaped our approach and outcomes. One significant hurdle arose during the data collection phase, where we faced usage limits from the BrightData API for navigating pages on Glassdoor, restricting the amount of data we could scrape. Also we had to study fundamental concepts of web scraping, exploring tools like Selenium and BeautifulSoup, which we hadn't previously mastered. As a result, we confronted numerous errors while scraping Glassdoor, so we had to overcome these obstacles.

Additionally, we grappled with computational limitations by the Azure server, preventing us from implementing algorithms reliant on heavy models. This constraint prompted us to explore more efficient algorithms that could deliver the desired results within the available computational resources. Moreover, restricted access to libraries and models on the server further challenged our implementation efforts, compelling us to seek alternative solutions and optimize our approach.

Furthermore, given the size of the project, we recognized the importance of devising a well-structured schedule to maximize collaboration efficiency and align efforts towards our shared goal. This necessitated careful planning and coordination to ensure productive teamwork and timely progress.

Moreover, the complexity of the provided LinkedIn datasets posed another obstacle, with many columns featuring nested subcolumns, requiring advanced data processing and management.

Reflecting on these limitations, we acknowledge their influence on our project's outcomes. Despite encountering constraints such as resource availability, technological limitations, and time restrictions, we leveraged these challenges as opportunities for growth and innovation. By embracing flexibility, problem-solving, and effective teamwork, we successfully navigated these limitations to deliver a robust solution.

Conclusions:

In conclusion, we have successfully achieved our objective of developing a personalized career pathing solution, integrating a diverse range of algorithms.

Our solution generated dynamic career paths tailored to LinkedIn profiles, aligning with their individual experiences and skillsets.

We firmly believe that our model holds significant potential for future applications, given its foundation in well-established machine learning algorithms, predictive models, and data analysis techniques.

Throughout this project, we explored various perspectives and disciplines, enriching our understanding of the interconnected fields that shaped our journey , and we saw that project planning is valuable for delivering high-quality outcomes.

Moreover, this project has provided invaluable experience in navigating real-world challenges as data scientists, fostering collaboration and effective management within a diverse working environment.