‘Resume Optimizer’ - Executive Summary
IoT Management, Spring 2024

Jeff Mathew Sam, Chris White,Josh Li, and Lansing Wilson

Introduction

Our project aims to utilize Databricks to analyze a dataset of 1.2 million current data science and data-analyst-related job listings extracted from LinkedIn Jobs. The goal is to optimize resume keywords by filtering for key terms, locations, salary ranges, and job titles, enabling us to identify relevant numerical and linguistic trends. To obtain these companies' operating industries, we had to merge our dataset with a 17 million global companies dataset based on location. Through the insights gained from our model, prospective data science applicants can expect a substantial improvement in their job application success rates.

Problem Statement

With the U.S. Bureau of Labor Statistics (BLS) projecting a nearly 35% increase in Data Science positions from 2022 to 2032, the competitive landscape in the job market is already emerging (BLS, 2022). While BLS data doesn't provide definitive statistics for professionals transitioning mid-career into data science or related tech fields, it's reasonable to assume that without a data science-related degree, the odds may not be in their favor. However, optimizing one's resume to align with automated algorithms and queries can empower individuals to compete and succeed in this competitive environment effectively.

Business Opportunity

With the projected increase of 35% in Data Science positions over the next decade, there's a remarkable opportunity to develop a model tailored for Data Scientists and related roles. What distinguishes our model from AI is its ability to provide a quantifiable dataset that can evolve and remain relevant to the current market. Moreover, as our model expands, it can branch out into other sectors. For instance, in the realm of business schools, general management positions without specific concentrations such as finance, accounting, strategy, and artificial intelligence are expected to grow by 10% in the next decade according to the BLS. Individuals opting for a concentration within these sectors can anticipate a growth rate of up to 25% in the next 10 years.

With the anticipated exponential growth in the number of individuals graduating with graduate or terminal degrees, there's a clear market for overall expansion and sustainability. The Resume Optimizer could boost success rates by an additional 5% for individuals seeking skills matching, gap analysis, and match-making recommendations. While this 5% increase isn't empirically proven, it's estimated based on the fact that roughly 73% of individuals find a position within six months of graduating from a graduate program.

By integrating the Optimizer, we could expect an 8% increase, bringing the match rate to approximately 83%. This figure aligns with commonly observed statistics among top university graduate programs. Essentially, this enhancement would be akin to hiring someone to tailor resumes to each job listing and submit them, thus significantly improving an individual's chances in the job market.

Model Explanation

The Resume Optimizer is a model built using the backbone of Python and SQL, utilizing a platform called Databricks to merge the two languages seamlessly. To test the model's rigidity and scalability, we chose to ask four fundamental questions from the data:

1. What U.S cities offer the most amount of opportunities for individuals seeking new employment at a Junior (Jr) or entry level or Senior (Sr) or more experienced level?
2. When applying for a position, what skills/keywords should be placed within your resume and LinkedIn profile to maximize your chances of securing a position?
3. If an individual has a dream company to work for, utilizing a keyword quarry, what is specific or recurring for that company?
4. Utilizing text analysis models, what keywords are most recurring and relevant when building your competitive resume to sync with an ATS algorithm built for your sector?

a. Data Pre-processing and data cleaning

Navigating a 5.1 GB file presented hurdles in the initial phase due to DataBricks' 2 GB upload limit. We overcame this by segmenting the file into 700-800 MB parts using AWS EC2 xlarge instance and Linux command line tools. Despite successful segmentation, accurate schema reading proved challenging due to encoding issues, resolved by exporting the file using Python's .to_csv function with "QuoteAll" and manually defining the schema in PySpark using StructField. Merging a 17 million-row industry dataset with LinkedIn data lacked a common key, necessitating the creation of one based on location, involving deciphering 28,000 unique locations. Ultimately, we merged both datasets, yielding 185,000 rows of data. Our data cleaning process included removing records with null values and unwanted characters from company names and job skills.

b. Models performance and outcomes

We have seen a great deal of success within our model and its performance. To identify industries within our dataset in the information below, we will answer the questions posed above with quantifying data and an assessment of the performance as it relates to the questions posed within the Model explanations. Our model successfully showed the top cities for people looking for Jr or senior-level positions.

For analyzing keywords from job_skills and job_summary we began by identifying job roles using n-gram: bigram text analysis, as job titles were missing. Splitting the dataset into 'Inexperienced or less experienced graduates' and 'experienced professionals', we employed Keyword Frequency analysis and Keyword Co-occurrence analysis models. Analysis of job_summary descriptions, along with Topic Modeling, revealed crucial insights. Notably, common keywords like 'business,' 'requirement,' and 'system' emerged for 'Business Analyst' in the 'IT service and consulting industry.' Similarly, 'Senior Manager' highlighted terms like 'business,' 'team,' and 'experience.' Leveraging these keywords from job_skills and job_summaries can enhance job seekers' visibility in the Applicant Tracking System (ATS).

c. Short comings & Limitations & Recommendations

Our model performed better than expected in all categories. However, there are a few aspects of the model that can be improved. To start, the size of the model as it relates to the data it uses is massive and robust. The size of the dataset almost made the model unworkable and unusable. To scale this model to expand into other countries not described within the original dataset would be quite challenging. Additionally, to use this model as a steady palace of the job market, we would need to utilize a more robust data processing system such as AWS or Microsoft Azure to keep up with the size and computing requirements of a global market.

In our analysis, biases arise from categorizing job prospects based on job title patterns, potentially overlooking relevant terms. Similarly, the selective removal of generic words and adjectives in job_summary analysis may skew keyword identification. Moreover, pre-processing steps like stop-word inclusion and custom exclusions introduce biases by potentially overlooking pertinent job_summary details. These biases should be considered to ensure the accuracy and fairness of our analytical outcomes.

Conclusion

As recommendations and directions for further research, we recommend incorporating data from other sources, such as Glassdoor, Indeed, and USAJobs, which would make the model's output even more robust. Additionally, we recommend including information on the job-applicant ratio for each job, current and future salary expectations for the job position, company culture, and values extracted from the company's website. This will enable job seekers to be more effective in their application process and increase their chances of landing more job interviews.

To scale we would need a much larger and more diversified data set that covers more industries and our dataset would need more than just senior jobs to cover the needs of potential users for resume optimizer to expand future competitiveness in the market.

References

1. https://www.bls.gov/ooh/math/data-scientists.htm
2. https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024
3. https://www.kaggle.com/datasets/mfrye0/bigpicture-company-dataset
4. https://grammarly.com
5. https://chat.openai.com/ - to correct grammatical errors, sentence structure