Milestone Report 3

LinkedIn Dataset 2023-2024 Analysis

Initial insights and trend identification

Group I

Beema Sajeev

Jaisy Joy

Parminder Kaur

Harsh Kundal

Diksha

Capstone-project/milestone-3.ipynb at main · Beema20/Capstone-project (github.com)

Table of Contents

Introduction

The goal of this project is to analyze the LinkedIn Job Postings dataset from 2023 to 2024 to extract insights into job market trends, salary predictions, and remote work dynamics. The project aims to provide valuable insights for job seekers, employers, and policymakers.

Key Tasks Completed

We analyzed the company distribution and identified that Macy's is the most active company in posting job listings, with approximately 3,500 postings. Other leading companies include Boldly Premium Executive Assistants, Hemphill – Search, Consulting, Staffing, and Millennium Recruiting, Inc. This high level of activity suggests a significant demand for new hires within these companies, likely reflecting their growth or turnover rates.

For industry distribution, we found that Industry ID 148.0 leads with over 25,000 job postings, followed by Industry IDs 284.0 and 156.0. This indicates that certain industries are experiencing higher growth or have higher hiring needs, which could be due to various factors such as economic growth, technological advancements, or seasonal hiring patterns.

In terms of job skills demand, job skill 22475 is in the highest demand, with nearly 30 postings requiring it. Other high-demand skills include 12608, 18185, and 5525. This highlights the importance of certain skills in the job market, suggesting areas where job seekers might focus their skill development efforts.

Additional Visualizations and Analysis

We created a scatter plot for salary distribution by industry, revealing a wide range of maximum salaries across different industries, with a few outliers indicating extremely high salaries. Most salaries are clustered at the lower end, suggesting that while there are high-paying jobs, they are not the norm across industries.

The correlation matrix for the postings dataset showed that variables such as views and applies are moderately correlated, suggesting that job postings with more views tend to receive more applications. Strong correlations between salary variables (max_salary_x, max_salary_y, min_salary_x, and min_salary_y) indicate consistent salary data entries. Weak or no significant correlations were observed for most other variables, indicating diverse factors influencing job postings.

Lastly, the distribution of company_size and employee_count showed a variety of company sizes and employee numbers, with larger companies tending to have more employees and followers. A strong positive correlation (0.73) between employee_count and follower_count suggests that larger companies attract more followers on LinkedIn.

Output Overview

Data Quality Issues Identified

- Inconsistent postings data entries which were rectified during preprocessing.
- Missing values in certain columns like remote_allowed which were imputed or analyzed for patterns.
- Outliers in salary data which were examined and handled appropriately.

Remediations

- Implemented robust data preprocessing steps to ensure data quality, including text preprocessing, data integration, categorical encoding, feature engineering, and numerical feature scaling.

Next Steps

- Conduct a deeper salary analysis by industry and job role to identify patterns and trends.
- Analyze the remote_allowed variable to understand the prevalence and trends in remote work opportunities.
- Perform a time series analysis of job postings to identify seasonal trends and hiring peaks.
- Compare the demand for specific job skills with their supply in the job market to identify skill gaps

Challenges Faced

- Data Integration: Merging datasets from different sources required careful handling to avoid data loss and ensure consistency.

- Handling Missing Data: Imputation strategies had to be carefully designed to maintain data integrity.
- Outlier Management: Detecting and managing outliers in salary and other numerical features posed significant challenges.

Conclusion

This milestone report highlights key trends in job postings, including the most active companies, industries with the highest demand for new hires, and the most sought-after job skills. The correlation analysis provides insights into the relationships between various job posting attributes, helping to understand the dynamics of job market activities. The EDA phase has laid a solid foundation for further analysis, enabling more detailed studies into salary distributions, remote work trends, and skill gaps.

This comprehensive analysis will provide actionable insights for job seekers to align their skill development with market demand, for employers to strategize their hiring processes, and for policymakers to understand job market dynamics and design informed policies.