Project 2: Topic Proposal

Econ 1680: Machine Learning, Text Analysis, and Economics

Lena Kim

March 30, 2024

1 Introduction

It is a known fact that many companies receive thousand and thousands of job applications a day, and as such rely on algorithms to screen resumes. In fact, a New York Times coverage of the ever-increasing reliance on resume algorithms has even led to the identification of keywords and skills to include in your resume to make it past this resume sifter (Weed (2021)). As a Post article states, including even one of these key words can mean the difference between getting hired and being discarded (Abril (2023).

Taking inspiration from this trend, I want to accurately predict the salary levels of a job based on keywords used in its job description, as well as the resumes that it gets. With some companies remaining tight-lipped in regards to salary transparency, this text analysis project may be useful for job seekers who want to know more about compensation before taking any steps toward subjecting themselves to a cold, algorithmic hiring process.

Research Question: Using keywords and skills in the job description, can we accurately predict the salary of the job?

2 Data

While I had originally planned to use a publicly available **Kaggle** dataset on job details and salary compensations of around 500 jobs, I will now use text scraping methods to get my data. Using an API from Adzuna, an open sourced site that includes updated job descriptions, salaries, and locations, I have scraped around the top 30 page results of jobs, and will create a dataframe of descriptions and salary as my main dataframe. The variables of interest will be:

- Job Title: The title of the job. (String)
- Key Skills: The key skills required for the job. (String)
- Industry: The industry of the job. (String)

- The salary of the job. (Float)

... and more or less as my research develops.

3 Method

In order to conduct a deep analysis of the skills and keywords at play, I will employ a number of methods in my text analysis application. I will first run a **sentiment analysis** on the variables defined above to see whether it can cleanly predict the salary of the job. Next, I will configure **clustering** and possibly **Support Vector Machines** to see if nonlinear methods will fit and predict the salary ranges well. In doing so, I will see which methods accurately predict the salaries, and which keywords and skills play the most weight with this.

4 Timeline

Task	Timing
Data download and cleaning	1-2 weeks. Feb 29
Descriptive statistics	2 weeks. Mar 8
Method development	2 weeks. Mar 22
Applying method	3 weeks. April 12
Formatting results and graphs	2 weeks. April 26
Drafting	1 week. May 3
Revising Draft	1 week. May 8

5 Generative AI Exercise

Copy and paste your text from the Introduction, Data, and Method section of your proposal into ChatGPT with the following prompt:

- `I am going to ask you questions about the following research proposal:` Insert Intro/Data/Method Text
- `For the described research project,` Insert the following questions

Here are the questions to ask ChatGPT about your research proposal. After each question, copy and paste the generated response.

1. "What is the research question of the topic proposal?"
   The research question for the described project is:

   "Using keywords and skills in the job description, can we accurately predict the salary of the job?"

2. "Why is this an economics research question?"
   While the primary focus of the research question appears to be on predicting job salaries using keywords and skills, it has economic implications for both job seekers and employers. Here are some reasons why this research question can be considered as having economic relevance:

   Labor Market Efficiency: The study aims to understand how well job salaries can be predicted based on specific keywords and skills. Efficient labor markets are essential for economic productivity and growth, and this research contributes to understanding the dynamics of the job market.

   Information Asymmetry: In many job markets, there is often information asymmetry between employers and job seekers regarding salary expectations. By predicting salaries based on job descriptions and skills, the research seeks to address this asymmetry, potentially leading to a more transparent job market.

   Wage Determinants: Understanding which keywords and skills are associated with higher or lower salaries can shed light on the factors influencing wage determination. This information can be valuable for both job seekers and employers in negotiating fair compensation.

   Labor Market Trends: Analyzing the data on job titles, key skills, and industries alongside salary information can provide insights into broader labor market trends. This information is crucial for policymakers, economists, and businesses in adapting to changing market conditions.

   Implications for Job Seekers: The research aims to benefit job seekers by providing them with insights into the salary expectations associated with specific keywords and skills. This economic information can empower individuals in making more informed decisions about their career paths.

   Productivity and Innovation: An efficient allocation of human resources, facilitated by a better understanding of salary determinants, can contribute to increased productivity and innovation within industries. This, in turn, has broader economic implications.

   In summary, while the research question involves text analysis and machine learning techniques, its focus on predicting job salaries based on keywords and skills makes it relevant to economics by addressing issues related to labor market efficiency, information asymmetry, wage determination, and broader economic trend

3. "What data source is the project using and does the student have access to it?"
   The research project is using a publicly available Kaggle dataset on job details and salary compensations of around 500 jobs. The dataset includes variables such as job title, key skills, industry, and the salary of the job. Since the dataset is mentioned as publicly available on Kaggle, it suggests that the student has access to it, as Kaggle datasets are typically accessible to the public for research and analysis purposes.

4. "What method does the project propose to use? Is it a text analysis or NLP method?"
   The proposed research project primarily suggests using text analysis methods, specifically sentiment analysis and possibly clustering and Support Vector Machines (SVMs),

to analyze the keywords and skills in job descriptions and predict the associated salary levels. While these methods utilize techniques commonly associated with natural language processing (NLP), the project's focus appears to be more on text analysis rather than NLP per se. However, NLP techniques may still be involved in preprocessing and feature extraction steps within the text analysis process.

References

Abril, Danielle (2023) "Your résumé isn't the only thing popular job sites evaluate," *Washington Post*.

Weed, Julie (2021) "Résumé-Writing Tips to Help You Get Past the A.I. Gatekeepers," *New York Times*.