Predicting and Recommending Skills for Job Candidates

Alex Tang
Eric Ge
Guru Balamurugan
Kyler Chen

1 Introduction

In an increasingly competitive job market, people seeking employment often face a complex landscape of required skills and qualifications. Employers frequently include a variety of requirements in their job postings, including both domain-specific technical abilities as well as more general interpersonal/analytical skills. Therefore, we see identifying and bridging the gap between a candidate's current skill set and the ideal one for a specific role as an important challenge. This report investigates the feasibility of utilizing machine learning methods to predict a job candidate's level of qualification as well as insight on skills they may be missing.

While recruiters, hiring managers, and workforce development agencies already use systems like this to understand the talent landscape better and identify potential candidates, our aim is for this technology to aid applicants themselves directly, by allowing them to identify areas of improvement within their portfolios. Furthermore, educational and training institutions can benefit from this as well, through tailoring their curriculum based on skill gaps in the market. The broad goal is to improve the career search processes of students by making skill requirements more transparent and actionable.

This project idea draws on existing work in skill extraction, job recommendation systems, and text mining of job descriptions. Although no external collaborator or professor specifically contributed to this particular work, we have been informed and inspired by prior research and open-source frameworks. The full code for this project, once completed, will be accessible at: `https://github.com/kylerc405/comp562-final-project`.

2 Problem and Motivation

The problem addressed in this project is to determine the missing skills a candidate should acquire to align themselves more closely with a given job opportunity. Unlike simpler classification tasks, this problem requires considering multiple sets of attributes: the candidate's current skill set, the job's commonly-seen skill set, and the description of the job. The output is not just a binary decision (e.g., suitable or not) but rather a ranked list of recommended skills for the candidate, as well as a numeric score as an index.

The motivation for this project stems from the growing complexity of the job market and the mismatch often seen between candidate skills and job requirements. When a job seeker begins applying for roles, we believe they likely would benefit from a straightforward, consolidated list of the skills/technologies they would benefit most from learning, based on real-world job data. Thus, they are given the opportunity to spend their time on the most valuable tasks for achieving their desired job field. Candidates who can identify and target specific missing skills stand a better chance of securing employment or progressing in their careers. The economic impact could be significant: improving candidate-job matching can reduce underemployment, support career transitions, and guide individuals toward more purposeful skill development.

3 Application Domain

The envisioned application involves both individual candidates and organizations that provide career guidance. For an individual user, the system would function as a personalized recommender: given their existing skill set and a target job's requirements, the model would output a set of skills that could increase their competitiveness for the role, as well as a score depicting their current level of desirability as a candidate. Such a tool could prove beneficial in improving their resume's efficacy in describing their abilities, as well as informing an applicant on whether or not a given job is worth their time to consider applying for. For organizations, such a tool could be integrated into career counseling platforms, learning management systems, or job portals, providing automated suggestions for skill enhancement pathways.

An example scenario might be a mid-career professional aiming to transition from a general administrative position into a business development role. Using their current skills (e.g., "Microsoft Excel", "Customer Service", "Scheduling"), the model examines the target role (e.g., requiring "Sales Strategy", "CRM Software", "B2B Relationship Management") and identifies missing competencies (e.g., familiarity with "Salesforce CRM"). The candidate could then focus their efforts on developing these targeted skills, either through online courses, workshops, or formal training, rather than aimlessly upskilling in random areas.

4 Related Work

In recent years, there has been significant research on skill extraction from resumes and job descriptions. Several projects have attempted to create skill-based recommendation systems for candidates. For example, the "Job Description Skills Extractor" extracts skills from job descriptions to match candidates. Additionally, the "Resume ATS Tracking LLM Project" focuses on using machine learning for resume ranking and skill extraction. Another notable project is the "AI Resume Analyzer", which uses natural language processing to analyze resumes. These projects are linked below in the references page.

These existing works inspired the design of our own skill gap analysis and recommendation system. Unlike the mentioned works, our approach provides a tailored skill ranking and analysis, that intends to help users target missing skills based on job training data, directly contributing to their career development.

5 Approach and Methodology

5.1 Data Loading and Preprocessing

The dataset used for this analysis was sourced from `https://www.kaggle.com/datasets/suriyaganesh/resume-dataset-structured?resource=download`, containing structured information extracted from professional resumes, normalized into multiple related tables. The primary goal was to perform a skills gap analysis using machine learning, aiming to predict the likelihood of an individual securing a computer science (CS)-related role based on their skills. This involved:

- Identifying key skills that distinguish successful CS professionals from others.
- Using model-derived feature importance to highlight skills gaps.

Preprocessing Steps:

- *Data Cleaning:* Imputation and normalization were applied to cleanse the data of NaNs/nulls and lexical inconsistencies such as commas and capitalization.
- *Date Processing:* Unstructured date formats were standardized to facilitate processing, ensuring that only the most recent work experiences and associated skills were considered for each resume.

5.2 Skill and Job Title Mapping

Mapping Strategy:

- Job titles were mapped against a predefined list of titles commonly associated with the CS field. Titles that matched were classified as CS professional roles.

5.3 Regression and Target Variable Definition

Instead of a binary classification of CS vs. Non-CS based on job titles, a continuous scoring mechanism was implemented:

- A score ranging from 0 to 100 was computed to quantify the aptitude of a candidate as a CS professional, factoring in weighted skills and educational qualifications such as degrees (e.g., Masters, Bachelors).
- This composite score was then used as a continuous target variable for the model, allowing for the ranking of individuals relative to one another based on their qualifications and skills.

5.4 Model Building and Processing

Feature Encoding and Dimensionality Reduction:

- Due to the presence of over 200,000 unique skills in the dataset, frequency filtering was applied to focus on skills that appeared more than 100 times in relation to CS-specific entries.
- Skills were further processed using a MultiLabelBinarizer for vectorization and Truncated SVD for dimensionality reduction, simplifying the feature space to the most relevant and common skills within the CS profession.

Model Training:

- A Random Forest classifier was trained using the processed features to predict the likelihood of candidates securing CS roles based on their resume data.

6 Evaluation and Impact

To evaluate the effectiveness of our approach, we conducted a series of tests with a sample dataset. We trained our model to predict the most important skills for a candidate securing a computer science (CS)-related role, based on the dataset. The key evaluation metrics included:

- **Accuracy**: We evaluated the model's accuracy on a test set that was 20% of our total dataset, to determine how accurate the model's predictions of skills.
- **Skill Importance**: We also analyzed the relative importance of skills using the trained Random Forest model, identifying which skills most influence a candidate's suitability for the role.
- **Missing Skill Recommendations**: Through gap analysis, we were able to identify the top missing skills that candidates should acquire to improve their chances of securing a CS role.

6.1 Results

- **Accuracy**: The model achieved an accuracy score of approximately 98% on the test data, demonstrating that our system can reliably predict the skills influencing the likelihood of obtaining a CS role.
- **Top Missing Skills**: Our gap analysis revealed key skills that candidates often lack. For example, skills like "RMAN," "Performance Tuning," and "DBA" emerged as critical missing skills for candidates applying to CS-related positions. Furthermore, the order of critical missing skills closely resembled that of the most important skills.

The impact of this project can be seen in multiple domains:

1. **For job candidates**: This tool allows candidates to clearly understand the skills they need to focus on in order to improve their employability for a specific role.
2. **For educational institutions**: By analyzing the skill gaps in the market, universities and training institutions can tailor their curricula to better prepare students for in-demand roles.
3. **For hiring organizations**: Companies can leverage this tool to better understand the talent pool and help candidates align their skills with market needs.

7 Conclusion and Future Work

This report focuses on the conceptual foundations, application domain and use cases, programmatic implementations. It proposes a machine learning-based system to analyze a candidate's skill set in relation to a target job's requirements and recommend additional skills for the candidate to acquire. The motivation for our project lies in improving job-market alignment, aiding candidates in strategic upskilling, and ultimately contributing to a more efficient and equitable hiring process.

Future work in this area could involve several enhancements:

- **Improved Natural Language Processing (NLP)**: We could enhance the ability to process unstructured data, such as job descriptions, by using advanced NLP techniques to extract skills more effectively from free text.
- **Bias Mitigation**: It's important to consider how biases in the training data (e.g., gender, race, or geographic location) might affect the recommendations. Future work could focus on ensuring that the model does not unintentionally disadvantage any particular group of candidates, by including more variables such as ethnic background or location.
- **Multi-domain Support**: While this model is tailored for CS roles, extending it to other industries (e.g., finance, marketing, healthcare, etc.) would provide broader applicability.
- **Feedback Mechanism**: Incorporating user feedback to fine-tune recommendations could enhance the personalization of the tool, allowing candidates to assess whether the recommended skills have helped them in real-world job applications.

By incorporating feedback and iterating on the model, we aim to further improve its recommendations and expand its scope to assist candidates in a wider range of job markets.

References

https://github.com/giterdun345/Job-Description-Skills-Extractor

https://github.com/Deba951/Resume-ATS-Tracking-LLM-Project

https://github.com/deepakpadhi986/AI-Resume-Analyzer