Table of Contents

Title 2

1.0 Background and Problem Description 3

1.1 Research Objectives 5

2.0 Research Methodology 6

Bibliography 7

Title

Consolidating Access to Candidate Data for Recruitment Headhunting: Leveraging Explainable Machine Learning

1.0 Background and Problem Description

To find candidates to headhunt for job positions, recruiters, defined as professionals who search and find qualified candidates for job positions (Jiawen, 2020), often have to manually search for candidates from various job portals. The recruiters individually visit multiple job portals such as LinkedIn, Pnet, Indeed and provide certain search criteria, such as job title, location, education, and other parameters. They then sift through the search results and manually save the candidates matching the provided search parameters to a tabular dataset format such as an Excel file for each job portal and further contact the candidates. This process is tedious and delays the process of finding suitable candidates quicker for the role (Pavithr & Gulomkodirova, 2024). The study by (Bogers & Mesut, 2021) explores how recruiters manually find candidates data across various job portals for headhunting, specifically focusing on their information-seeking behaviour and engagements with the job portals. Nonetheless, this study does not suggest how recruiters can unify access to candidate data from various job portals. According to (Peicheva, 2022), recruiters use Applicant Tracking Systems (ATS) to manage incoming job applications, however the functionality of these systems is not meant to consolidate candidate data profiles from various job portals when recruiters are searching for and headhunting candidates. In light of this gap in current research on manual recruitment process for searching and headhunting candidates, this study seeks to investigate various methods that may effectively consolidate candidate information from several job portals. The aim is to simplify the task for recruiters having to visit various job portals to manually obtain candidate information for job positions they are headhunting for. This study aims to do this by utilizing one of the existing professional data aggregation Application Programming Interfaces (APIs) such as Coresignal, an API that provides publicly accessible data about employees such as their experience history, skills, education, certifications, country from various job portals worldwide (Coresignal, 2025) or Proxycurl, an API that provides access to enriched candidate data from multiple job portals (Proxycurl, 2025). There are other similar APIs that provide access to such candidate data but this research will explore the usage of one of them based on wider data source coverage. Upon consolidating candidate data from one of the data aggregation APIs, this research will also explore the use of machine learning recommendation engines to match and rank candidates to job descriptions in order to further enhance the recruitment process of finding the most suitable candidates to headhunt. The consolidated data obtained from the mentioned APIs is presented in a JavaScript Object Notation (JSON) format which is not immediately suitable for directly training machine learning algorithms, therefore this research will also explore an effective data processing and transformation method to make the data ready for training machine learning recommendation engines. This additional layer of analysis through machine learning recommendation system aims to further improve the process of recruiters identifying the most ideal candidates for the position from the consolidated candidates by ranking and matching each candidate's profile against the job description using information like their work experience, education, skills, and certifications. Rather than solely ranking and matching candidates to the job description, this study will incorporate explainability to offer clear explanations for why certain candidates rank higher than others for the role, providing transparency to recruiters. The importance of building machine learning algorithms that are not only accurate but also explainable to recruiters is underscored by (Beretta, et al., 2024) who emphasize the need for transparency to build trust and ensure fairness in AI-driven recruitment. Studies in existing literature either only report on accuracy based on cosine similarity without explanations into the factors influencing the candidate rankings or use the SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) evaluation metrics for explainability. For example, a study by (Aram Khasro, et al., 2025) demonstrated high accuracy rates using SVM and LSTM models achieving 94% and 92% accuracy in matching and ranking candidates, respectively, but did not incorporate explainability in order to understand why certain candidates were a better match than others. Another study by (Magham, 2024) explores the role of explainable machine learning in AI-driven recruitment using the SHAP and LIME evaluation metrics. While SHAP and LIME are established tools for machine learning algorithm explainability, their technical nature can be complex and not easily understandable by non-technical users (Xianlong, 2024). This research will employ Shapash, a python library dedicated to the importance and interpretability of features in a machine learning model (Read the Docs, 2025). This library provides summary reports and visualizations of model feature importance with clear human-readable explanations (Read the Docs, 2025), making it easier for non-technical users such as recruiters to interpret.

1.1 Research Objectives

The primary aim of this study is to develop a solution that effectively tackles the difficulties of manually accessing candidate data from various job portals for headhunting in recruitment and integrate machine learning recommendation engines to rank and match candidates with explainability. The proposed objectives are as follows:

- Develop a methodology that utilizes professional data aggregation APIs such as Coresignal or Proxycurl. These APIs provide access to consolidated publicly available employee data from multiple job portals such as LinkedIn, Glassdoor, GitHub and Indeed. These professional APIs operate within legal frameworks for data collection, ensuring compliance with data privacy regulations, which is an advantage over manual scraping or unofficial data collection methods (Coresignal, 2025). While such APIs typically require paid subscriptions, they offer free trials with sufficient data access for research purposes, making them viable for this study.
- Develop a pipeline to transform the data obtained from the API, which is provided in unstructured data formats that are not immediately ready for machine learning uses. This includes converting the data into a tabular dataset format that enables the application of machine learning algorithms for candidate ranking and matching.
- Apply machine learning recommendation engines to the consolidated and processed candidate data to rank and match the candidate profiles from the search results against the job descriptions. This evaluates the candidate profiles according to how well they align with particular job descriptions based on their information such as work experience history, education, skills and certifications. This added layer of analysis seeks to improve the precision of candidate-job matching and facilitate recruitment agencies in swiftly identifying the most relevant candidates for the role.
- Implement explainability in the machine learning model using Shapash. This will involve creating interpretable models and visualizations that highlight the key features contributing to a candidate's ranking and matching for a role.

2.0 Research Methodology

To achieve the objectives of this research, the following methodology will be adhered to. The proposed methodology is intended to be adaptable across any industry in the recruitment domain:

- This research will evaluate and select from available professional data aggregation APIs such as Coresignal or Proxycurl based on comprehensive coverage of multiple sources of job portals. An API with broader coverage of multiple job portals will better address the multi-platform manual search of candidates. For research purposes, the free trial options available from these providers will be used as they offer exactly the same publicly available candidates data offered by paid subscriptions.
- Transform the unstructured data obtained from the professional data aggregation API into a structured tabular format suitable for machine learning use.
- Create a pipeline that addresses data issues such as incorrect data formatting, missing values and duplicates in order to clean and normalize the candidate data.
- Extract relevant features from textual data such as skills, experience history, education and certifications for machine learning.
- Develop a machine learning recommendation model that computes cosine similarity scores between candidate information and job descriptions. The trained model will return candidates ranked by similarity scores. Higher ranking candidates will be the ones most recommended for headhunting because they better fit the job description.
- Provide explainable candidate rankings from the trained model using Shapash's human interpretable reports and visualizations to highlight why certain candidates were ranked higher or lower to a particular description.

Bibliography

Pavithr, M. & Gulomkodirova, M. S. Q., 2024. A Study on Various E-Recruitment Tools and Its Effectiveness for Recruitment. *Kokand University Herald,* Issue 10, p. 33.

Coresignal, 2025. *Employee API: Unlock millions of professional profiles.* [Online]
Available at: https://coresignal.com/solutions/employee-data-api/
[Accessed 27 March 2025].

Bogers, T. & Mesut, K., 2021. *An Exploration of the Information Seeking Behavior of Recruiters.* Amsterdam, s.n., p. 3.

Jiawen, L., 2020. Recruitment and Selection. *Financial Forum,* 9(3), p. 1.
Peicheva, M., 2022. Data analysis from the applicant tracking system. *HR & Technologies,* Issue 2, pp. 6-15.

Proxycurl, 2025. *About.* [Online]
Available at: https://nubela.co/proxycurl/about
[Accessed 05 April 2025].

Magham, R. K., 2024. Mitigating Bias in AI-Driven Recruitment : The Role of Explainable Machine Learning (XAI). 10(5), pp. 461-469.

Beretta, A., Ercoli , G., Alfonso, F. & Riccardo, G., 2024. *Requirements of eXplainable AI in Algorithmic Hiring.* Amsterdam, Netherlands, s.n.

Aram Khasro, J., Abdulhady, A., Joan, L. & Mohammed Noori, E., 2025. Machine Learning for Recruitment: Analyzing Job-Matching Algorithms. *TechRxiv,* p. 1.

Read the Docs, 2025. *Shapash.* [Online]
Available at: https://shapash.readthedocs.io/en/latest/overview.html
[Accessed 07 April 2025].

Xianlong, Z., 2024. Enhancing the Interpretability of SHAPValues Using Large Language Models. p. 1.