Comprehensive Report on AI-Powered Résumé Screening/Fraud Detection

Team: Tahreeq | INNOV8-2.0 Finals, Eightfold.AI, IIT Delhi

1. Introduction

The problem we aimed to solve involved evaluating a database of 1,000 résumés to identify fraudulent entries and extract meaningful insights from the data. Our goal was to develop a presentable and easy-to-use dashboard for HR professionals to analyze the authenticity of candidates' claims and make informed hiring decisions.

This comprehensive report aims to explain our approach and methodologies, including data analysis, feature extraction, and results obtained from our AI-based tools and models.

2. Objective and Problem Statement

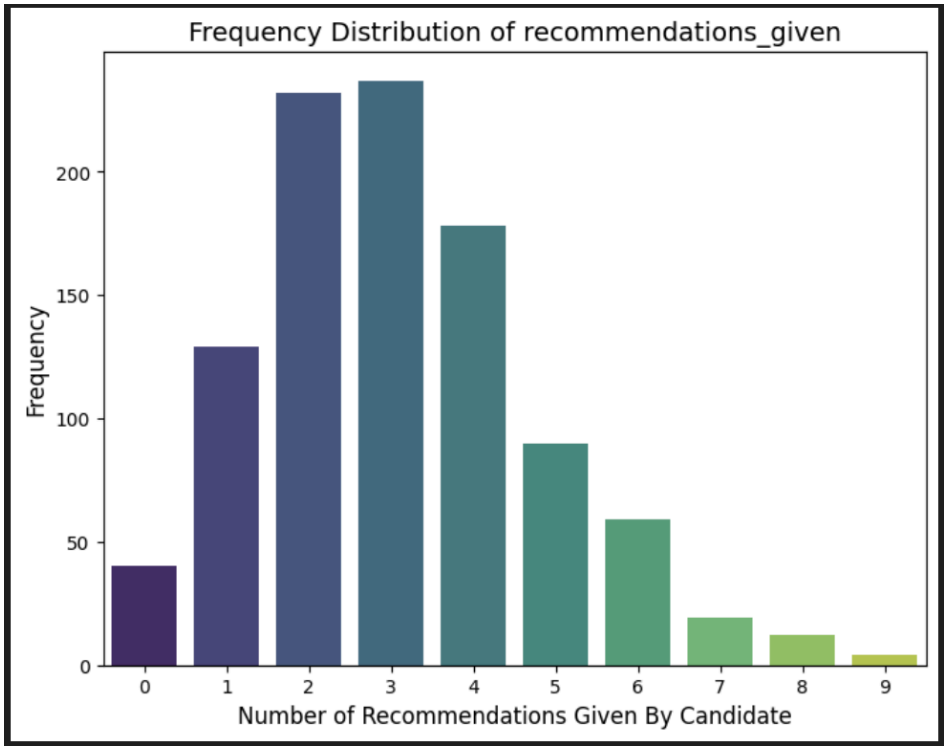
Our tasks include:

- Finding résumés where candidates had lied or falsified their experience.
- Extracting meaningful insights from the data.
- Creating a dashboard for HR professionals to present the data and insights in an accessible way.

3. Data Analysis

We started by analyzing the 1,000 résumé dataset. After cleaning, 964 valid entries remained. The first step was to explore the distribution of recommendations and endorsements made by candidates, which provided insight into patterns in the data.

- Number of Recommendations: Most candidates had recommended 1–5 other candidates. We plotted this distribution to identify outliers and abnormal cases.



This analysis helped in identifying suspicious résumé patterns based on the number of recommendations made and received.

4. Methodology

4.1 Text Extraction

We employed ***PyMuPDF*** for extracting text from PDF résumés. This allowed us to streamline the data pipeline and ensure that we could handle résumés in different formats consistently.

4.2 AI-based Feature Generation

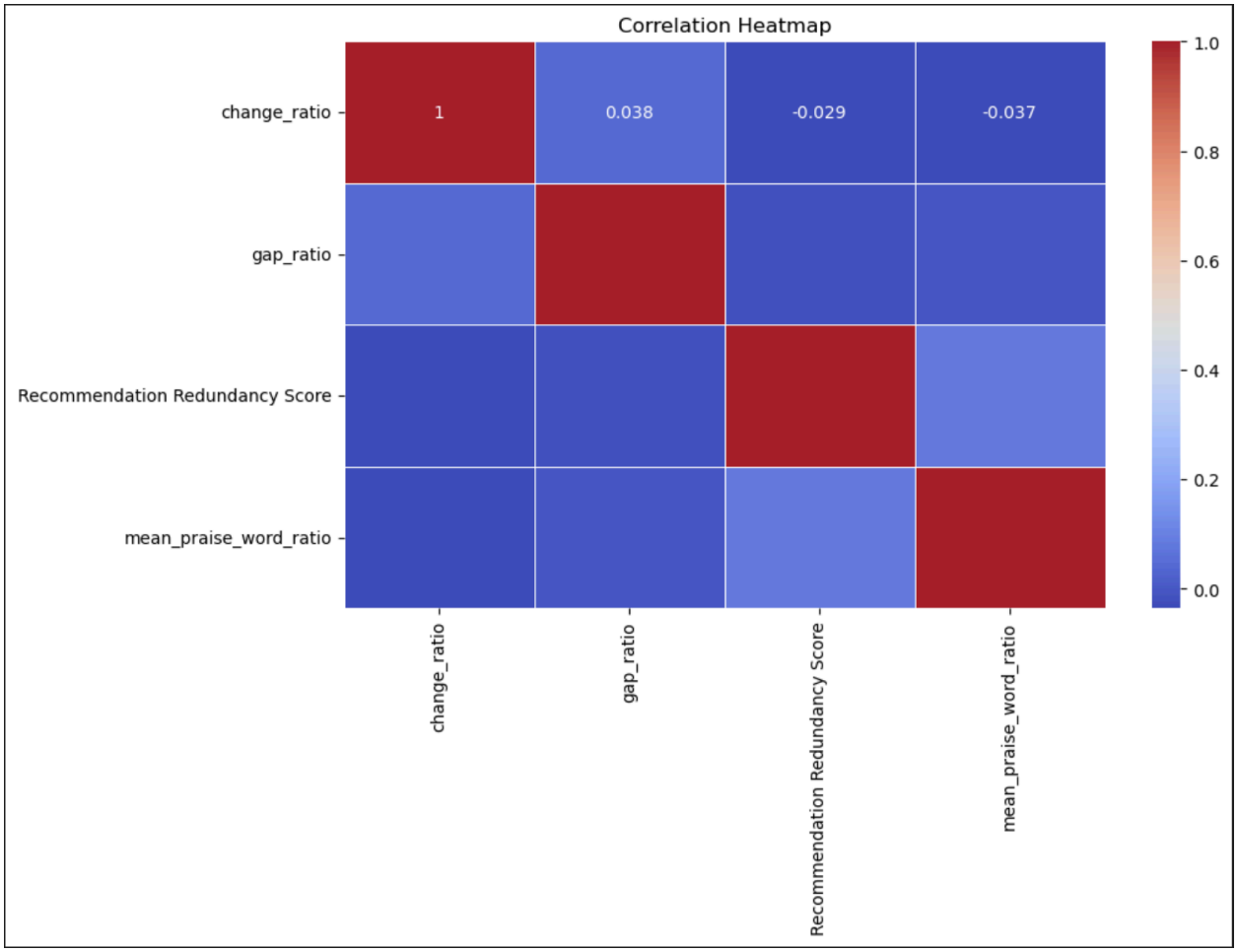
Using ***LLaMA-3.1-70b*** and the ***Groq API***, we generated key résumé features. Our goal was to assess résumé integrity by extracting details such as work experience and employment gaps.

1. **Circular Endorsements**:

   - We used ***NetworkX Graphs*** to calculate circular endorsements, identifying patterns of candidates endorsing each other in cycles of size 2–5. Smaller cycles (e.g., size 2 or 3) are more likely to be fraudulent, as they suggest collusion between candidates.

2. **Résumé Summaries**:

   - A crafted prompt in LLaMA enabled us to generate summaries of each résumé, from which we extracted key features using ***regular expressions***.



4.3 Experience Calculation

- We extracted work experience details from the summaries and calculated the duration of each candidate's experience. Additionally, we identified gaps in their careers, as significant gaps often indicate red flags in job applications.

Key Metrics:

- **Mean Experience**: We calculated the average work experience and career gap across all candidates, which served as a baseline for relative comparisons.
- **Total Number of Jobs**: The total number of jobs extracted helped assess résumé completeness and accuracy.

5. Fraud Detection via Redundancy Scores

5.1 Cycle Detection

We focused on detecting closed endorsement cycles (circular endorsements) between candidates. These cycles were weighed based on their size, with smaller cycles (size 2 and 3) being considered more suspicious.

Redundancy Score Calculation:
- Cycles of sizes 2, 3, 4, and 5 were assigned weights based on their prevalence. The redundancy score was calculated by summing the weighted cycle counts. This score helped identify potentially falsified recommendations or endorsements.

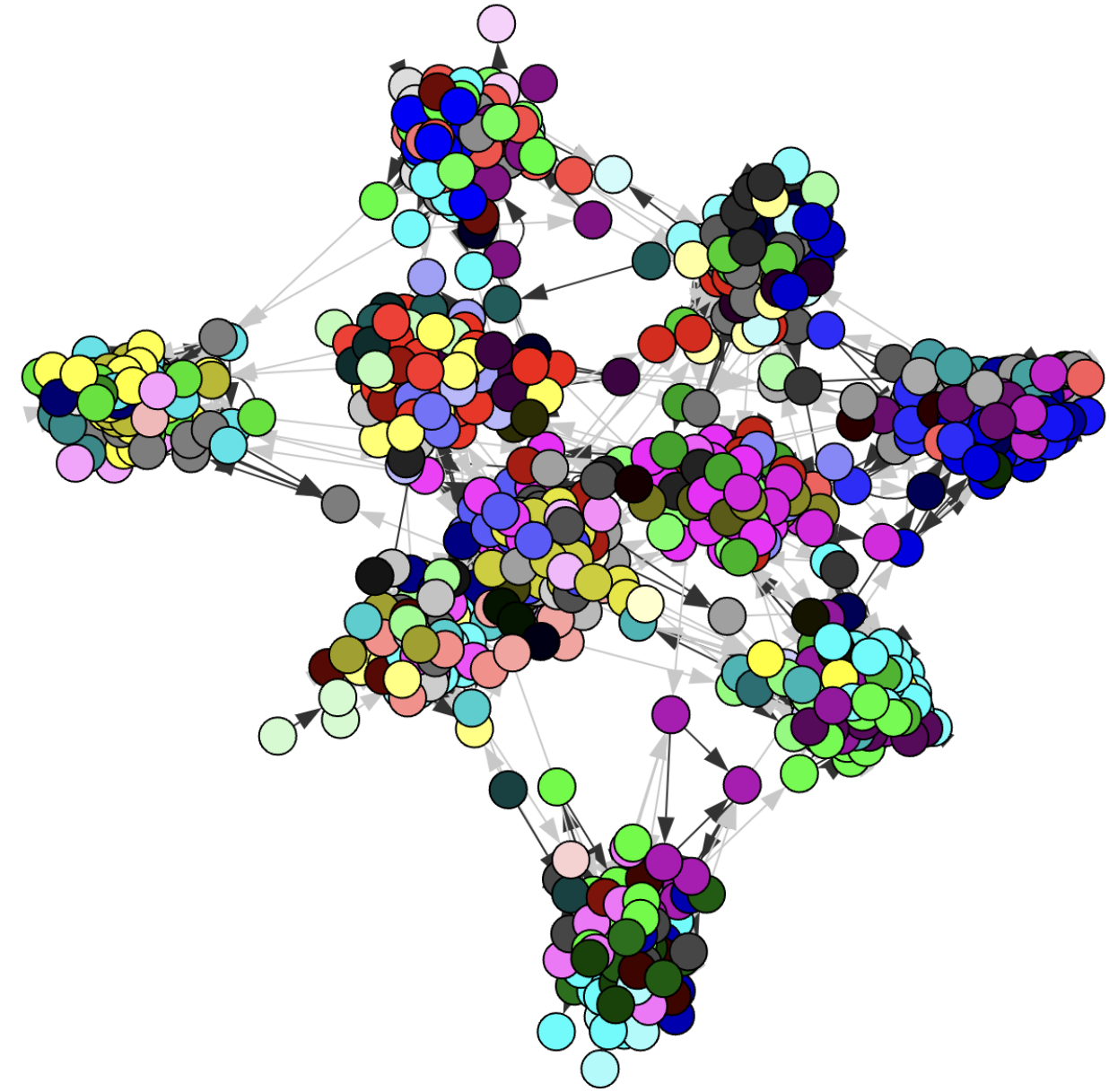
**[Insert graph showcasing redundancy score distribution]**

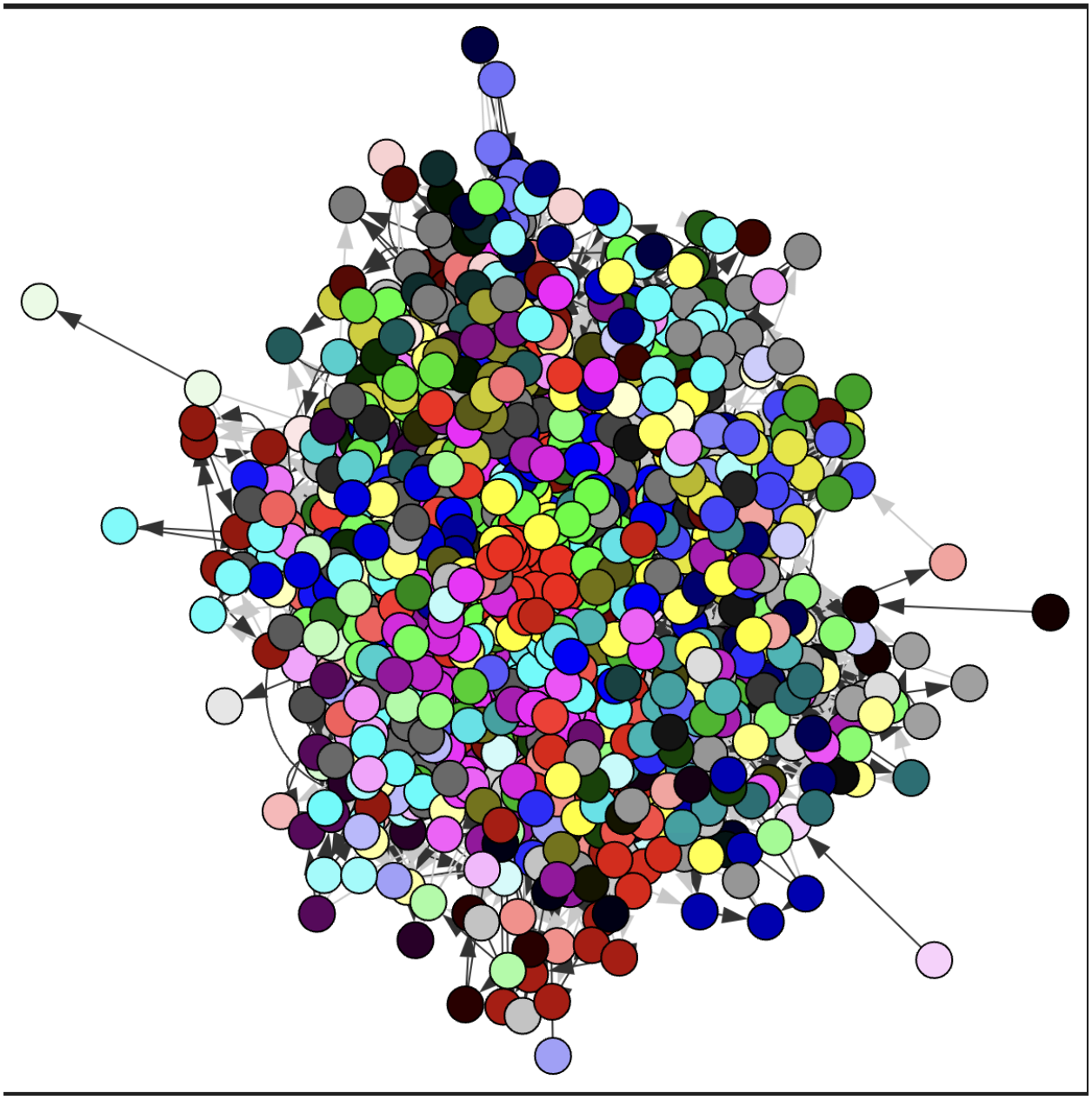
6. Community Detection and Clustering

We utilized the ***Leiden Algorithm*** with ***NetworkX Graphs*** to detect communities within the data. Community detection was applied twice:

1. **Initial Clustering**:
- Clustering was first performed without considering education, identifying groups of candidates who endorsed each other frequently.

2. **Education-based Clustering**:
- We incorporated the education feature into the community detection process to further refine the clusters. This approach helped differentiate between legitimate endorsements (e.g., within the same educational institution) and potentially fraudulent ones.



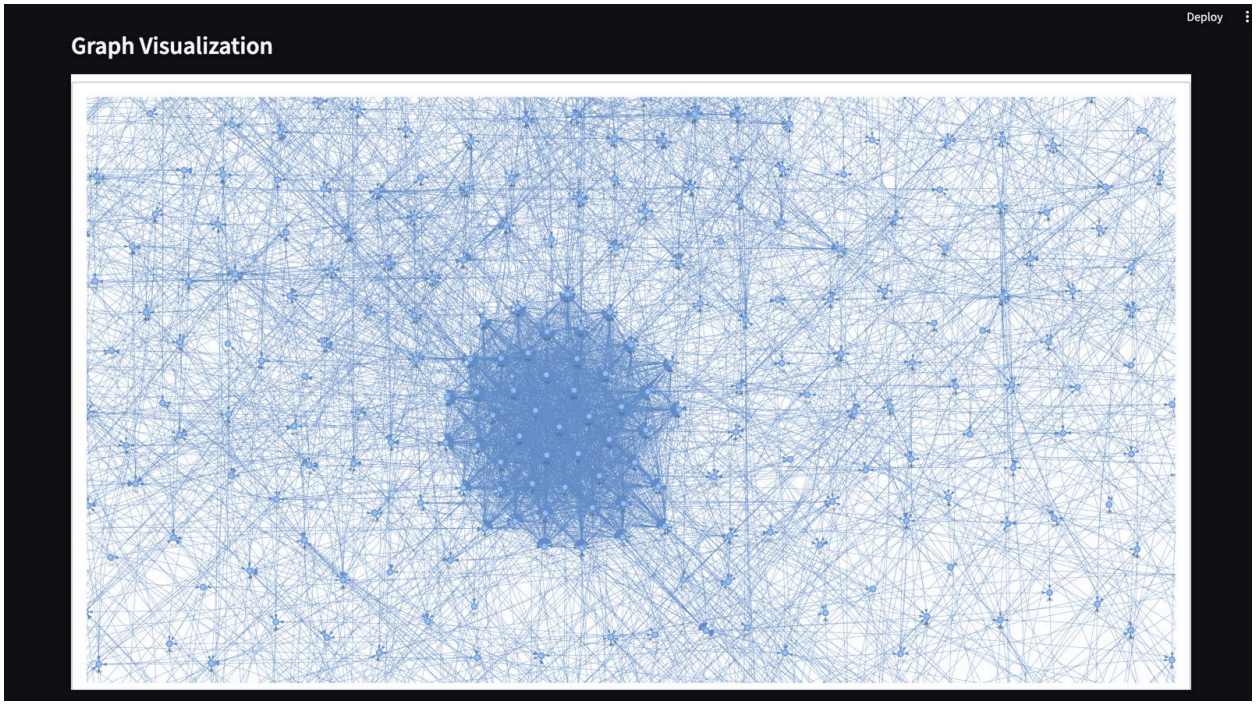


7. Dashboard Development

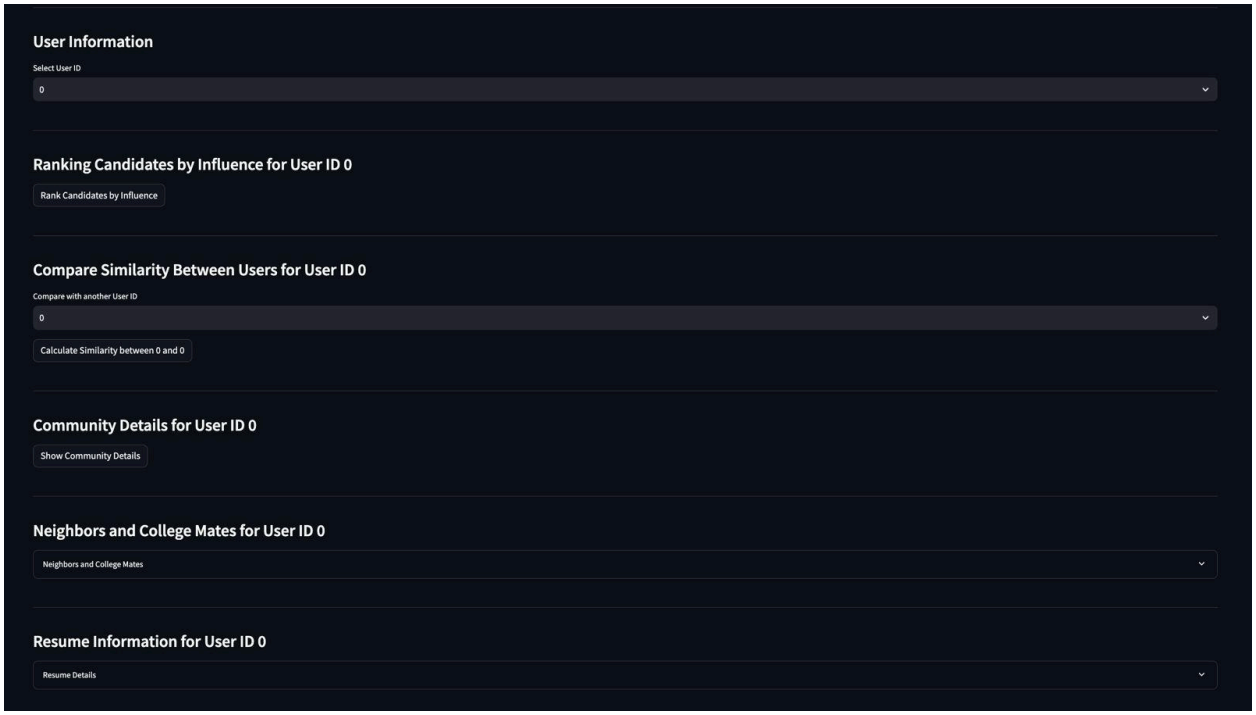
We created an intuitive and user-friendly dashboard for HR professionals to access the analysis. The dashboard provides key insights such as:

- Résumé scores.
- Fraud detection highlights.
- Community and clustering information.

- Graphical representations of endorsements and recommendations.



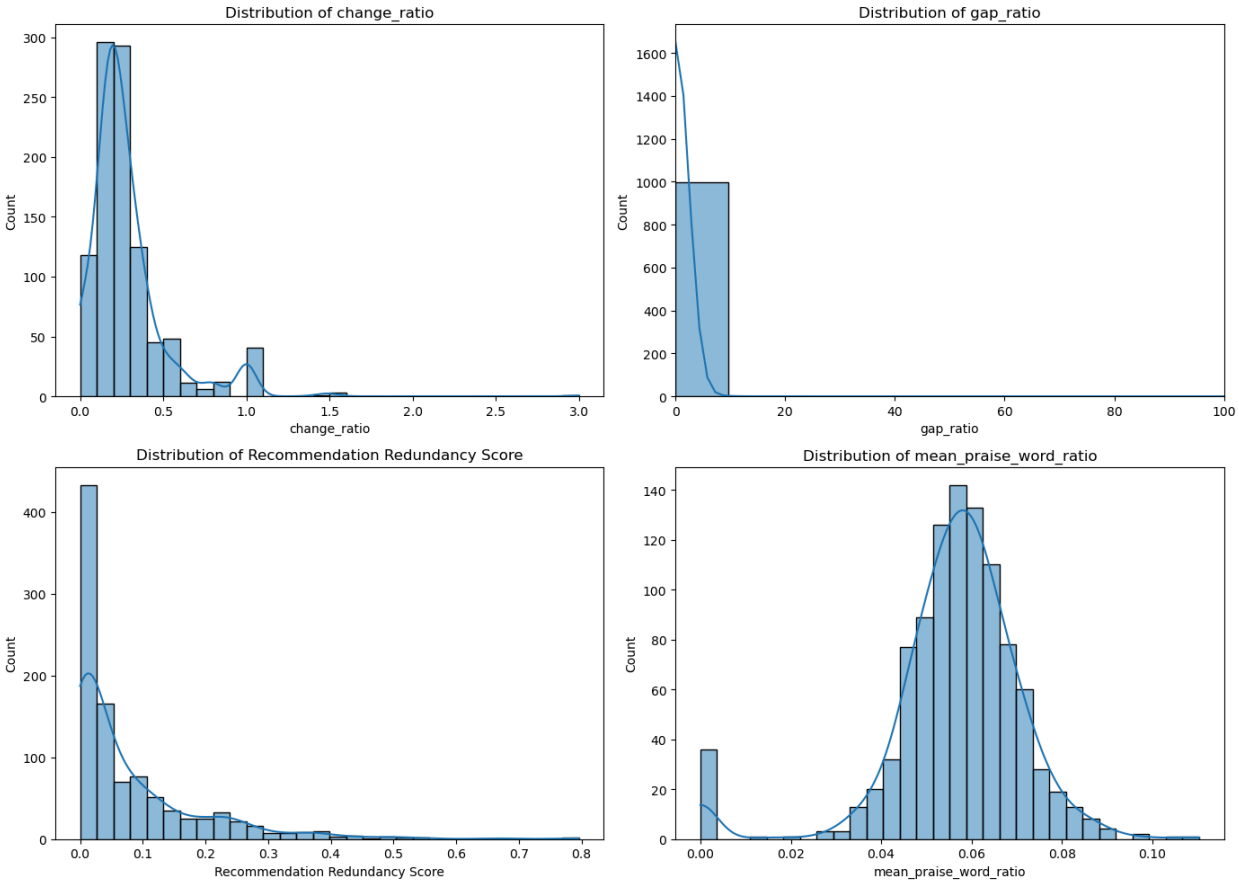
The dashboard ensures HR professionals can easily navigate through the data and identify potential red flags without needing technical expertise.

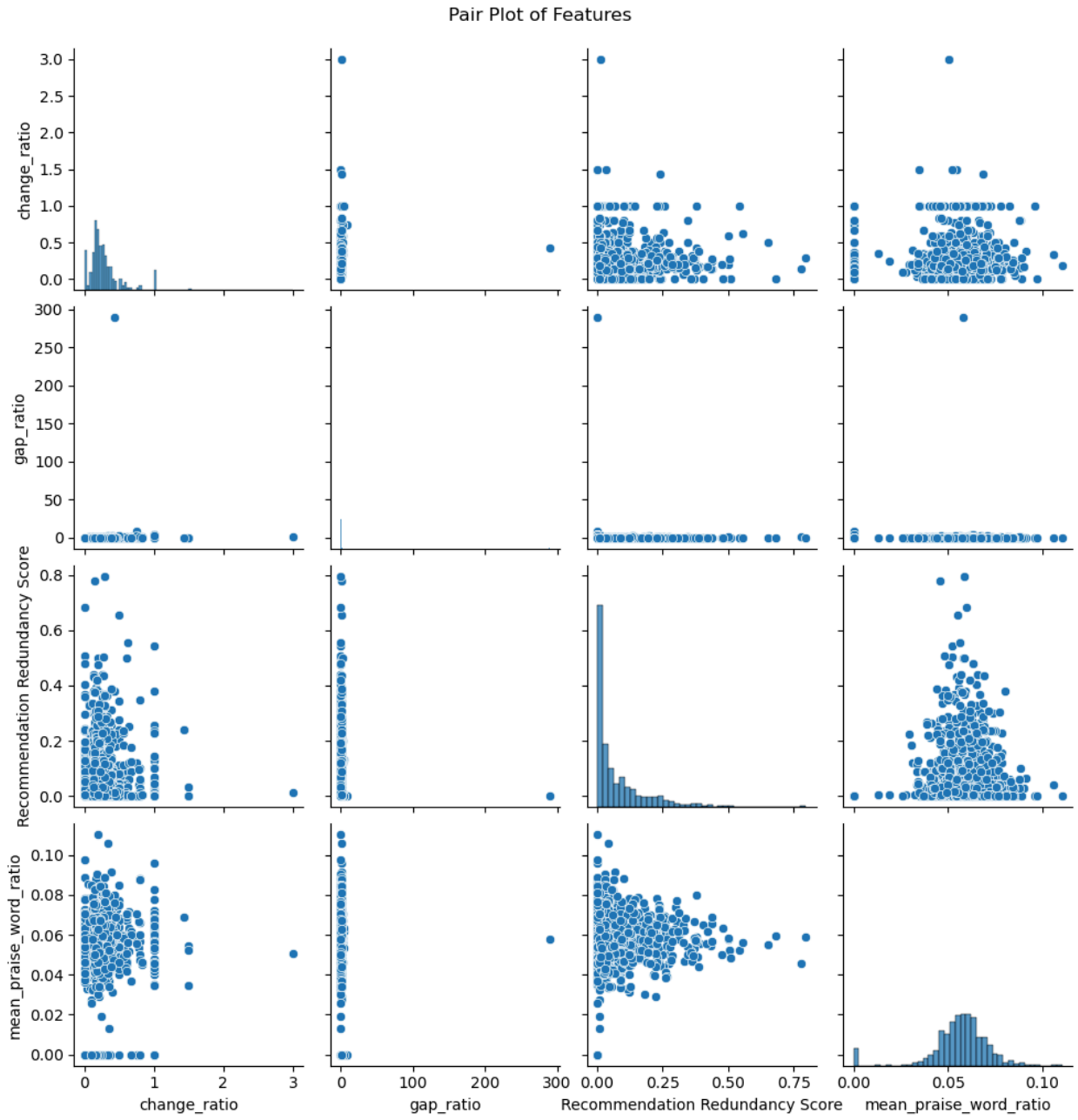


8. Results and Insights

Key Findings:

- Several résumés displayed circular endorsement patterns, indicating possible collusion.
- Candidates with large gaps in employment or inconsistent work experience durations were flagged as higher risk for falsified claims.
- The comprehensive scoring system effectively identified suspicious résumés and provided HR with actionable insights.
- LLM based scoring doesn't perform as good as extensive feature extraction like GAP in employment, number of changes made over the employment period and redundancy of recommendations
- We employed community detection and graph enhancement using education as a feature to better understand connections





9. Conclusion and Future Work

Our approach effectively identified fraudulent résumés through a combination of feature extraction, fraud detection, and dashboard visualization. We employed LLM based scoring as well but found that manual feature extracting and looking for patterns in the data is much better an approach. We also employed community detection to better understand the connection within candidates. in the future, we aim to enhance the model by incorporating the graph based connections to communicate the

probability of fraud using sophisticated NLP techniques like neo4j graph creation etc. and expanding the dataset for better generalization.

Additionally, we hope that more nuanced résumé features, such as skills matching and industry-specific benchmarks, to further improve fraud detection and résumé evaluation accuracy.
