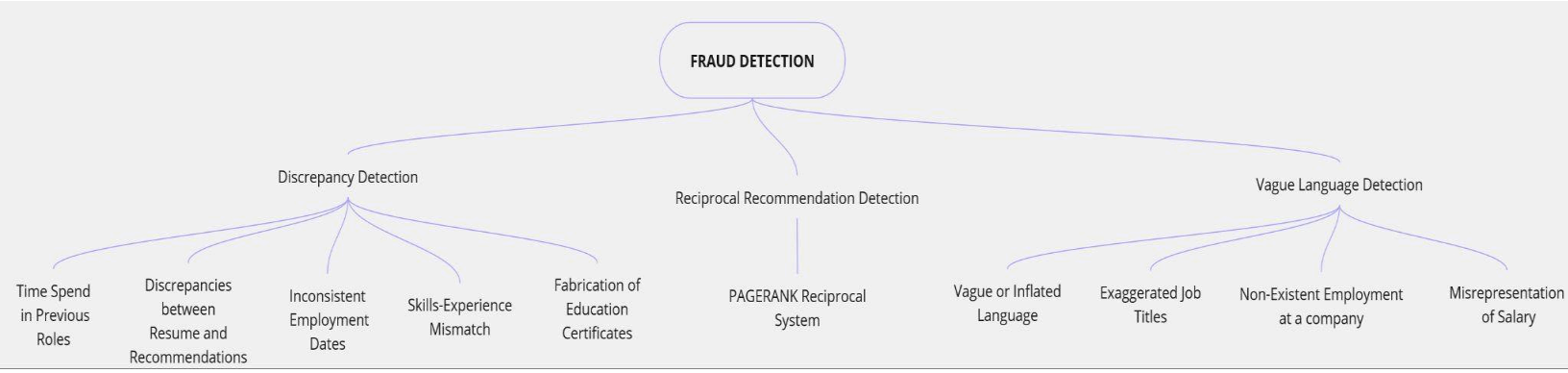
Fraud Detection Model

Our first order of business was to decide on the metrics to test each resume on. We formulated some potential Red Flags or indicators that if found in a resume would assign it a quantitative score that would translate it to being a Fraud resume. A resume will be tested on multiple of these metrics and a score is assigned for each metric, all of these scores at the end will be accounted and accumulated to tell if a resume is fraud or not.
We finally formulated 10 of these metrics divided into three categories visualized as:-



Description and Quantitative methods to measure for each of the metric are as follows:-

1. Discrepancy Detection

a. Time Spend in Previous Roles

Feature: We looked for unusually short tenures in past jobs, which may suggest job-hopping or embellishing experience.
Quantitative Measurement: We could calculate the average job duration and compare it with industry benchmarks.

b. Discrepancies between Resume and Recommendations

Feature: We compare resume information (e.g., job roles, achievements) with the content of recommendation letters.
Quantitative Measurement: We used cosine similarity to assess alignment between resume and recommendation letters. Low similarity may indicate fabrication or exaggeration.

c. Inconsistent Employment Dates

Feature: We identified gaps or overlaps in employment history that are not justified in the resume.

Quantitative Measurement: We calculated the **duration** of each job role and checked for overlaps or unreasonable gaps using simple rule-based systems.

d. Skills-Experience Mismatch
Feature: Identify cases where the listed skills don't match the claimed job experience or projects.
Quantitative Measurement: Use skills-job title matching algorithms to verify if the candidate's listed skills align with the expectations of their previous job roles.

e. Fabrication of Education Certificates
Feature: This is a potential Red Flag that could be implemented. It checks if there is any fabrication in the Education Certificates in the resume.
Quantitative Measurement: We could cross-check educational claims via databases like National Student Clearinghouse or certification bodies.

2. Reciprocal Recommendation System

a. PAGERANK Reciprocal System
Feature: It detects circular or reciprocal endorsements, where two or more individuals repeatedly recommend each other in different contexts to inflate credibility.
Quantitative Measurement: We constructed a **graph network** where nodes represent individuals and edges represent endorsements. We used **PageRank** to detect unusually dense clusters or loops of recommendations.
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. We use this same logic but conversely, as when there are too many connections or recommendations especially of circular kind between resumes, they are flagged for fraud.

3. Vague Language Detection

a. Vague or Inflated Language
Feature: We identify the use of generic or vague phrases that don't indicate specific achievements or roles (e.g., "great potential," "good team player").

Quantitative Measurement: We perform sentiment analysis and keyword analysis to flag overly positive but unspecific language. Additionally, evaluate the frequency of vague adjectives using Natural Language Processing (NLP) tools.

b. Exaggerated Job Titles
   Feature: Some candidates may inflate their job titles to appear more senior or responsible (e.g., claiming to be a "Director" when they were a "Manager").
   Quantitative Measurement: We could use word embeddings to evaluate if the title claimed matches the responsibilities described in the resume.

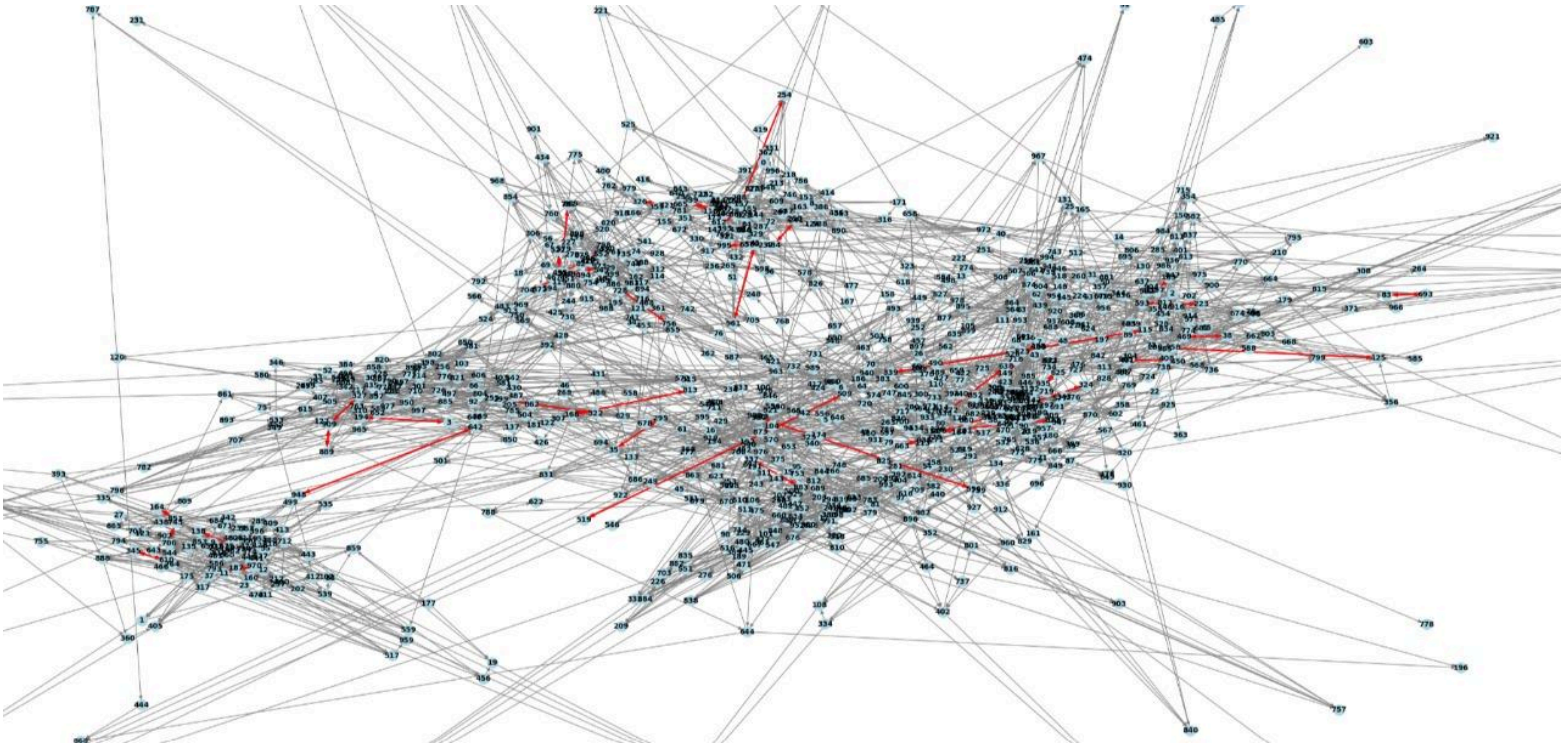
c. Non-Existent Employment at a Company
   Feature: This is another potential metric which concerns some candidates who may list employment at a well-known company where they never worked.
   Quantitative Measurement: We could use background verification services to confirm employment, like **LinkedIn** or other social media to cross-check employment claims.

d. Misrepresentation of Salary
   Feature: This is also a potential metric which concerns candidates who may lie about their salary to negotiate a higher package or exaggerate their responsibilities.
   Quantitative Measurement: We could compare the claimed salary with industry standards for the candidate's experience and location by using platforms like Glassdoor or Payscale.

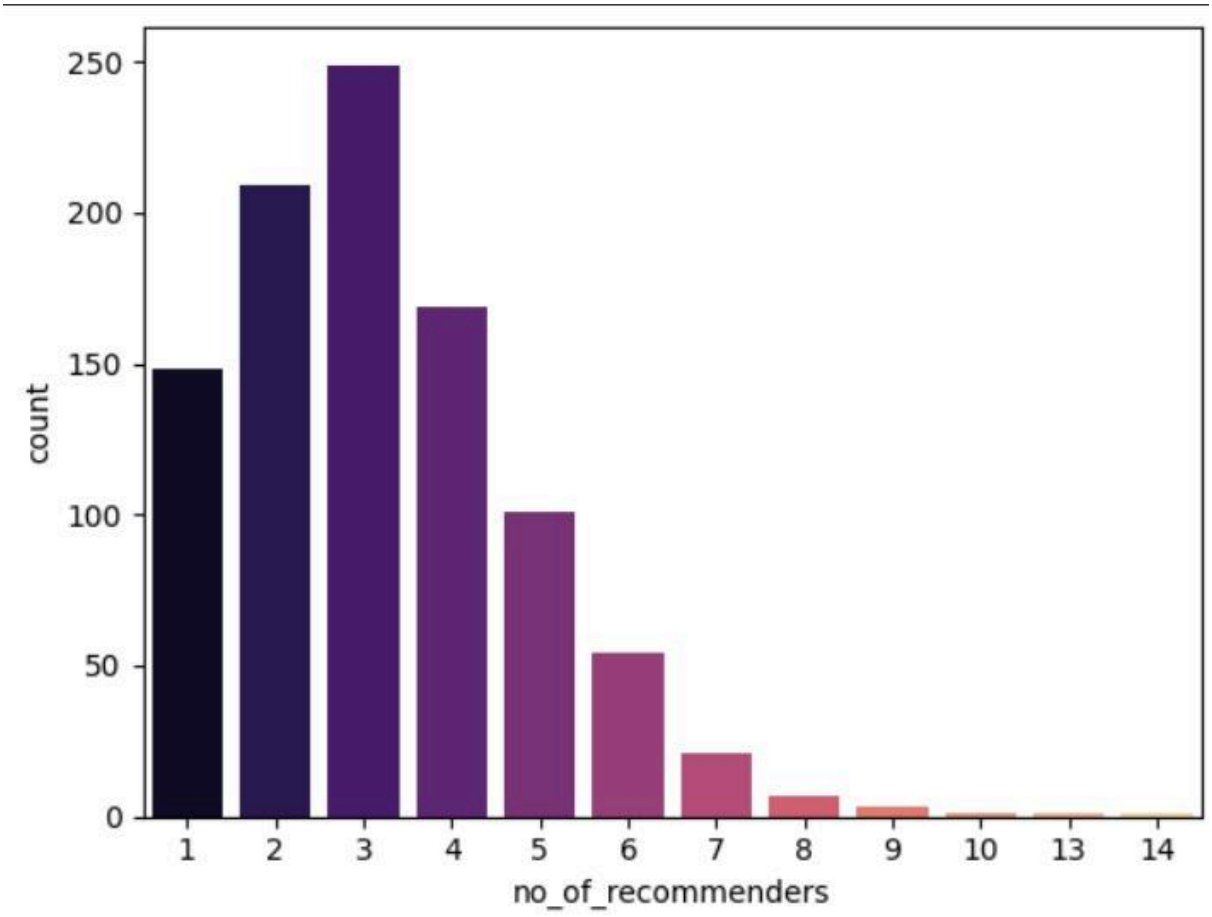
Key Findings of Fraud Detection Model

RECIPROCAL CONNECTION

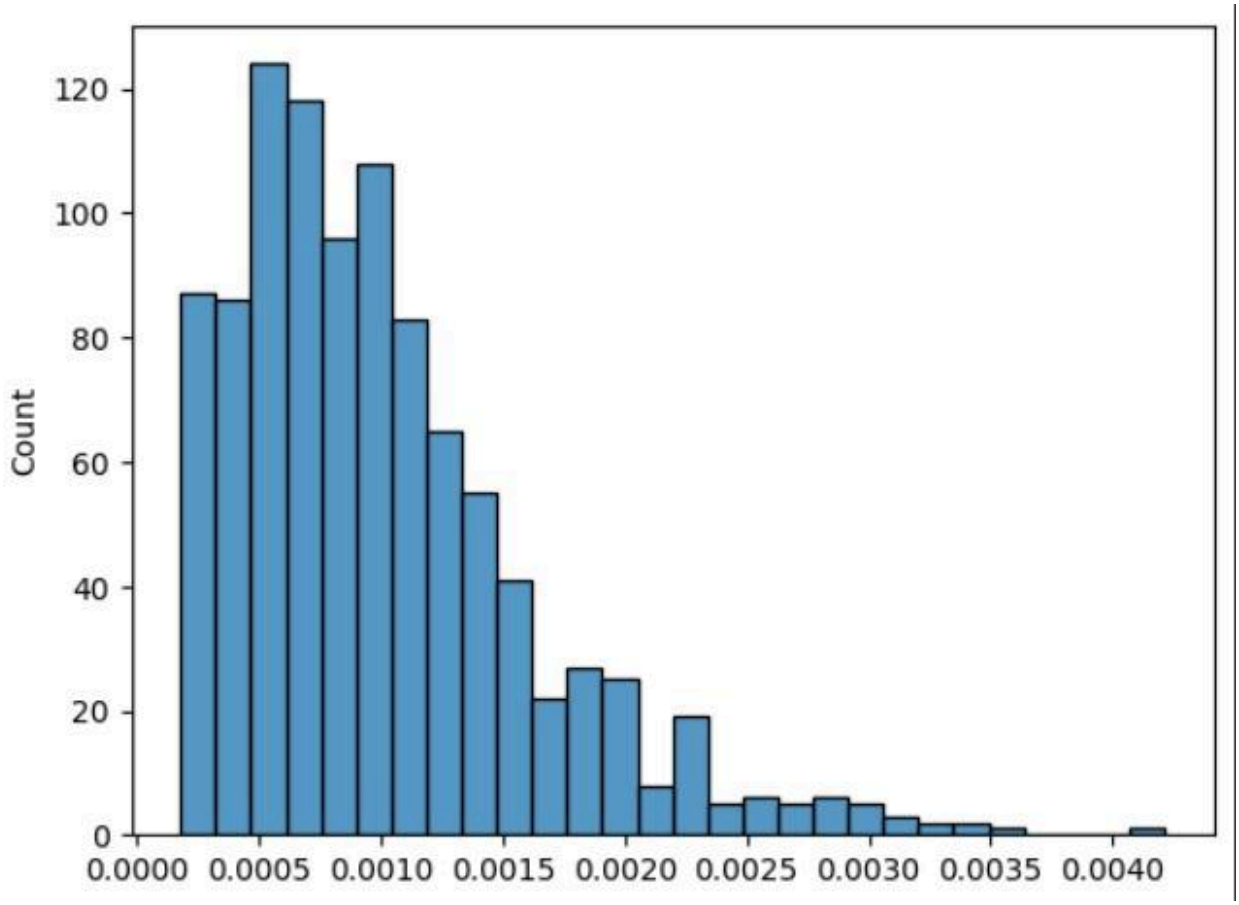
Using Pagerank we could plot the graph that would show connections between different candidates, it would look like this, the red lines show connections that are reciprocal or circular in nature.



Now, the graph that illustrates count of number recommendations is:-

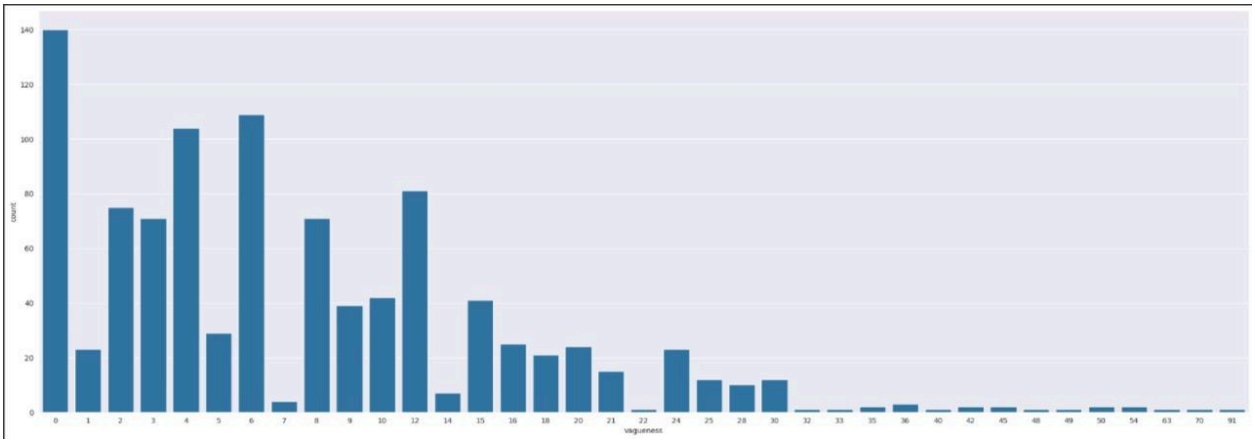


We concluded that the individuals who have 7 or more recommendations are high risk candidates, assigning them more pagerank score, which also looks like this (x axis in this graph is page rank score):-

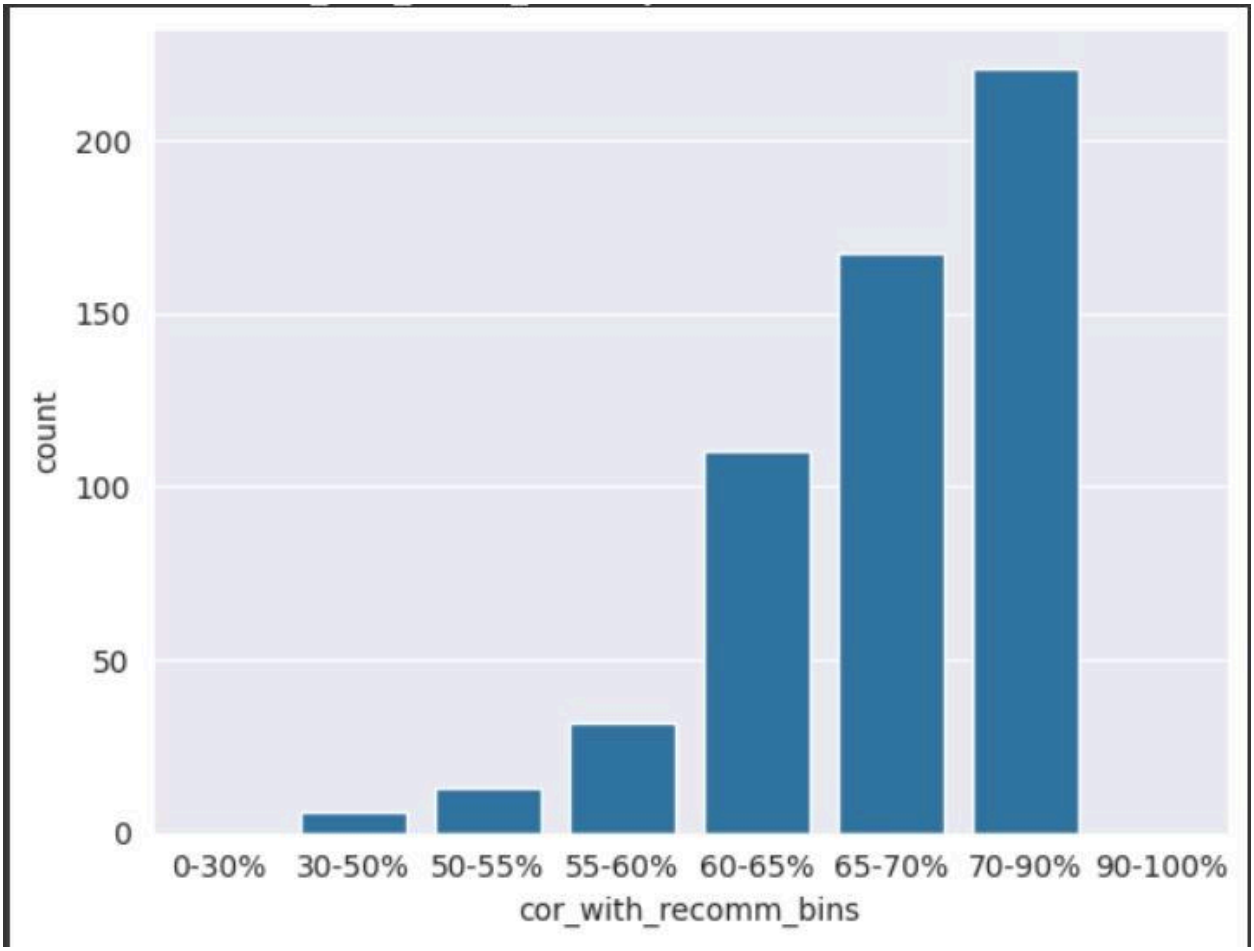


Vagueness

The graph of count vs number of vague or inflated words in recommendation letters for each candidate is:-



This shows that most of the recommendations have either no or low number of vague words. The recommendations that have a very high number of vague or inflated words (>20) are high risk candidates.
Lastly, the graph that shows correlation between the various recommendations with their respective resumes. As we can see, most of the data is highly correlated with the resumes, but some of the data has low correlation, which can be attributed to random writing of recommendations and not being properly in contact with the applicant, leading to the content of the recommendation not matching with the content of the resume. We can assume this low correlation resumes to be more likely to be fraud resumes.



DATA ANALYSIS

For a deeper understanding of the resume dataset and recommendations, we created a keyword list which shows strength and quality of these connections. We went through each resume's recommendation letters, opened the file and read all the lines into a list, then we iterated through each sentence in the document to find those quality keywords or phrases and count occurrences.

We created a metric named “meaningful connection score” by dividing the quality keywords and phrases by total sentences. Resume with less meaningful connection score tends to be a fraud, which we cross checked with our fraud detection model and those with higher scores have good content in their recommendation letters.

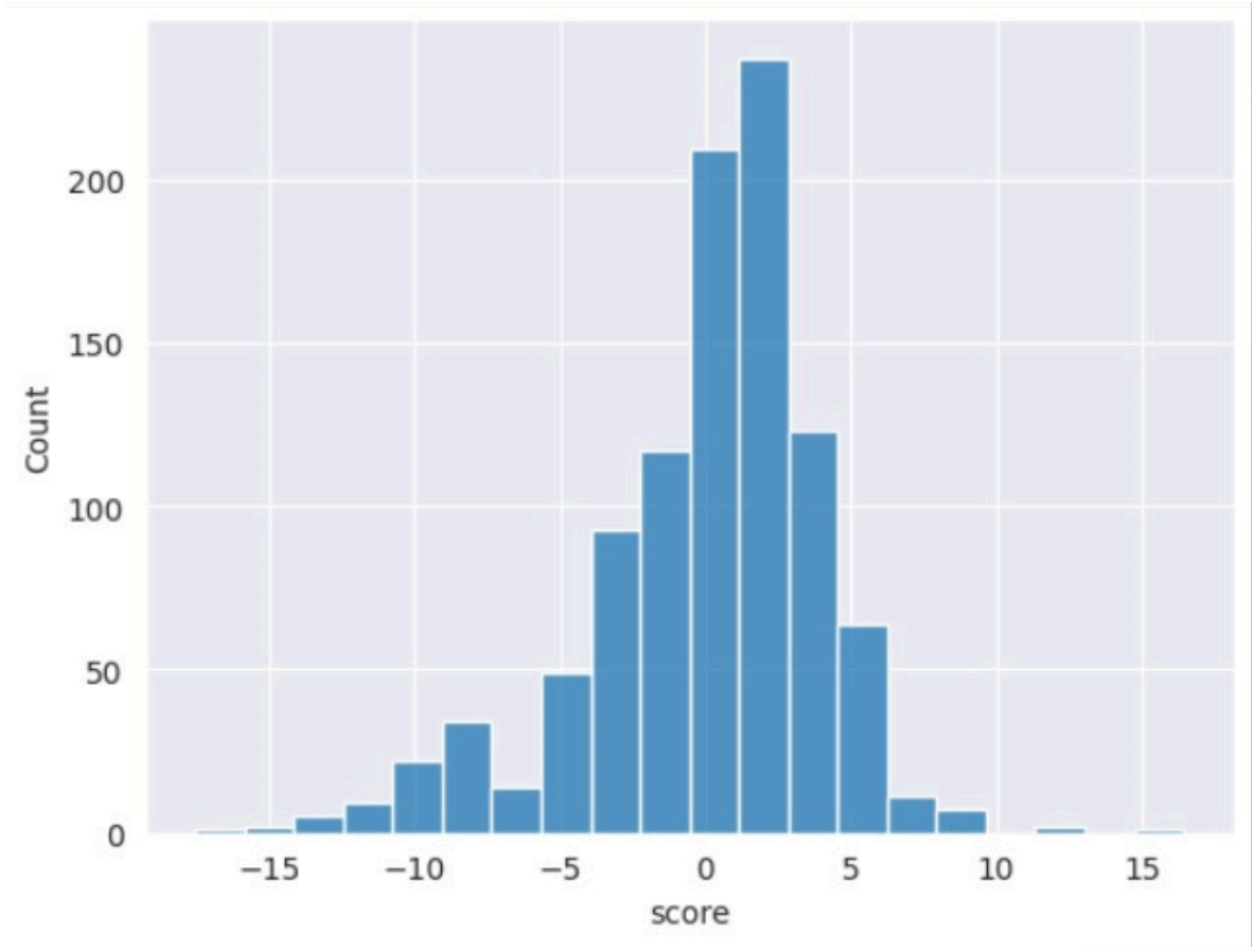
We further extracted the Skills stated by candidates in their resumes, and added it as a list to the dataset. Using this we found inter skill similarity, as well as similarity between the job stated and the skills provided.

Job	Average_Skill_Similarity	0
bilingual language arts sixth grade teacher	0.853286	[arts, English, instruction, Latin, letters, L...
mortgage banking default operations specialist ii	0.879560	[Adobe, Photoshop, streamline, Attorney, audio...
guest lecturer	0.883845	[basic, Council, English, Instructor, LANGUAGE...
accountant	0.870890	[Accounting, approach, AS400, auditing, bank r...
staff accountant	0.872650	[Expert in customer relations]

We used Word2Vec Embeddings to find the relation between the job names and the skills given by the candidate.

We also found the inter skill similarity so as to ensure that the skills given are consistent with each other and there is no sharp difference in skills given.

Finally the numerical based columns were transformed such that a higher score means a better resume, while a lower score means a poorer resume and a higher chance of it being a fraud. Weights were assigned to the transformed columns and added to give a final score dataframe. The scores in this ranged from -17.5 to 16.5.



A score less than -10 has a high chance of being a fraud while a majority of the resumes are near 0, which is they are neither very well made nor frauds, which is to be expected.

SCALABILITY

The system is designed to handle datasets beyond 1,000 candidates, using efficient algorithms for both fraud detection and data analysis.

- **Graph Processing**: We employed **NetworkX** for graph construction and centrality calculations. The time complexity for centrality measures scales as O(n2), where n is the number of nodes (individuals). For larger datasets, **distributed graph processing** tools like **GraphX** could be implemented to handle millions of nodes.
- **NLP Efficiency**: By using **TF-IDF** and **pre-trained embeddings (Word2Vec, BERT)**, we reduced the processing time for each résumé and recommendation letter. These models are capable of running in parallel, allowing the system to scale across large datasets efficiently

OPTIMIZATIONS

Early Stopping in Community Detection: For very large datasets, community detection was sped up by introducing **early stopping criteria** based on modularity score convergence. This reduced computation time without compromising the quality of community detection.

Batch Processing for NLP Tasks: We implemented **batch processing** for textual analysis tasks (e.g., cosine similarity, sentiment analysis) using **multiprocessing** to minimize latency.