ML (1).docx

by Aditi Nayak

Submission date: 04-Aug-2024 10:04PM (UTC+0530)
Submission ID: 2427086678
File name: ML_1_.docx (45.29K)
Word count: 5804
Character count: 35696

CA1: Think-Pair-Share

Team Members:
Aditi Nayak
Barish Chetia
Jil Kapadia
Ishaan Bhadrike

Topic Name: Machine Learning Job Listings on Glassdoor

Reference No.	Methodology Used	Domain	Data Set Used	Performance	Limitations	Outcome
[1]	Adapted CRISP-DM for structured data mining. Collected data from Glassdoor using Selenium. Conducted data preparation and preprocessing. Employed multi-class classification for job position prediction. Tested 19 ML algorithms, selecting the top 5 for a heterogeneous ensemble model.	Job position classification	The dataset contains 955 unique cases, including job positions such as data scientist, data engineer, analyst, machine learning, manager, director, and other job positions.	The heterogeneous ensemble models achieved an accuracy of approximately 100% using soft voting.	Limited by a sample size of 955 instances. Future studies should use larger datasets for better algorithm comparison.	The paper primarily measures the classification of job positions based on variables including average salary, company size, revenue, job description, and company rating.
[2]	Content analysis of online job advertisements for AI and ML	Artificial intelligence (AI) and machine	The data set used in this study was online	The study indicates that educational	The findings may be used to improve educational	AI roles prioritize deep learning,

	positions posted on Indeed.com Ranking of relevant skills for AI and ML positions Pairwise comparison of skill requirements between AI and ML positions	learning (ML) job advertisements.	job advertisements posted on the Indeed.com website.	programs and hiring practices should adapt to these skill demands, though the focus on a single source (Indeed.com) and pairwise comparison limit the comprehensiveness and representativeness of the findings.	programs and hiring practices, implying that the current state of these may be a limitation. The scope of the study was limited to a pairwise comparison between AI and ML positions, which may not capture the full range of relevant positions. The use of job advertisements from a single source (Indeed.com) may limit the representativeness of the sample.	NLP, and computer vision, while ML roles emphasize statistical modeling, data preprocessing, and algorithm development.
[3]	Formulating the job recommendation problem as a supervised machine learning problem Exploiting past job transitions and data associated with employees and institutions to	Job recommendation	The data set used in this study consists of job transition information extracted from publicly available employee	Ioannis K. Paparrizos, B. B. Cambazoglu, and A. Gionis (2011) evaluated the job recommendation system's	The limitations include potential biases from publicly available profiles and the model's dependency on historical data, which may not reflect future	The primary outcome measured in this study is the accuracy of predicting an employee's next job transition using a machine

	predict future job transitions Training a machine learning model on a large dataset of job transitions extracted from publicly available employee profiles		profiles on the web.	performance through experiments that revealed the machine learning model could accurately predict job transitions, significantly outperforming a baseline that always predicted the most frequent institution in the data.	job market trends.	learning model.
[4]	Evaluating studies that considered the temporal and reciprocal aspects of job recommendations Analyzing the fairness of algorithms used in job recommender systems Classifying hybrid job recommender system models using existing recommender system taxonomies	The domain of the paper by Radhika Taneja and Dr. Ashima Mehta (2023) is job recommender systems (JRS).	Utilized datasets from job recommender systems, focusing on temporal, reciprocal, and fairness aspects in job recommendations.	1) The ability of JRS models to provide improved recommendations by considering temporal and reciprocal aspects of job recommendations. 2) The ability of JRS models to provide a more balanced distribution of applicants	Potential limitations in model performance and balanced distribution of applicants in current JRS Limited literature on fairness of algorithms in JRS, with current approaches being insufficient Lack of consideration for the generalizabilit	The study highlighted the improvement in job recommendations by considering temporal and reciprocal aspects, enhanced applicant distribution, but identified limitations in fairness, bias, and generalizability of current job recommender systems.

				across similar job types. 3) The fairness and lack of bias in JRS algorithms, which is an important but often overlooked aspect of performance.	y of JRS across different datasets	
[5]	1) Operationalizing organizational culture (OC) as a word vector representation using job descriptors. 2) Validating this OC construct using language from 650,000 Glassdoor reviews 3) Applying the OC construct to Glassdoor reviews to quantify OC by sector 4) Validating the OC measure on a dataset of 341 employees and showing it explains job performance	Organizational culture	1. 650,000 Glassdoor reviews, which were used to validate the researchers' operationalization of organizational culture as a word vector representation. 2. A dataset of 341 employees, which was used to validate the researchers' measure of organizational culture and its relationship to job	Job performance of the employees.	The study may not have directly addressed interventions or tools for improving employee functioning, and further research in this area would be valuable. The sample size used to validate the measure of organizational culture was relatively small (341 employees). Quantifying organizational culture is inherently challenging	The study operationalized organizational culture using job descriptors, validated it with 650,000 Glassdoor reviews, applied it to quantify OC by sector, and demonstrated that the OC measure explains job performance using a dataset of 341 employees, despite its inherent complexity and

			performance.		due to its subjective and complex nature.	subjective nature.
[6]	The research uses a combination of machine learning (ML) and deep learning algorithms to classify fake job postings. The process involves data collection, pre-processing (lowercasing, removing nulls, tokenization, punctuation removal), feature extraction using semantic analysis and natural language processing, and model training and evaluation	The domain is online recruitment and job posting analysis, focusing on identifying and preventing fake job postings through the application of advanced ML and deep learning techniques	The dataset used for this research consists of nearly 18,000 rows of job postings, with 17 columns containing textual and numerical data. The data was sourced from Kaggle and the University of Aegon	The study leverages several ML algorithms, including XGBoost, for high accuracy in classification. The exact performance metrics are not specified but the use of advanced algorithms like XGBoost suggests an emphasis on achieving high classification accuracy	The paper does not explicitly discuss limitations, but common challenges likely include handling imbalanced datasets, the complexity of feature extraction, and ensuring the generalizability of the models to diverse job postings.	The research demonstrates the efficacy of combining ML and deep learning techniques in detecting fake job postings, highlighting the potential for these methods to enhance online recruitment processes by preventing data theft and cybercrime.
[7]	The paper utilizes a hybrid recommendation system combining collaborative filtering and content-based filtering. Natural language processing (NLP) techniques, including cosine similarity, are employed to match student skills with job requirements.	This research focuses on the domain of job recommendation systems within the broader context of online recruitment. It aims to enhance job recommendations for	The dataset comprises job postings and candidate profiles, sourced from various online job portals. It includes text data about job description	The system's performance is evaluated using precision, recall, and F1 score metrics. These metrics assess how well the recommendations match user preferences	The paper acknowledges challenges such as handling the dynamic and temporal nature of job openings and managing sensitive personal information. Additionally, the system's reliance on the	The proposed system successfully enhances job recommendations by leveraging NLP and machine learning. It provides personalized job suggestions that align

	The system integrates machine learning algorithms to improve recommendation accuracy	students by analyzing their resumes and job listings using NLP and machine learning	s, skills, and qualifications, which are pre-processed for training the recommendation system	and job requirements, indicating the system's accuracy and effectiveness in generating relevant job suggestions	quality and completeness of input data can affect its recommendation accuracy	with students' skills and preferences, thereby improving the efficiency of the job search process for both job seekers and employers
[8]	The study employs a hybrid recommendation algorithm tailored to dynamic user profiles. It updates and extends profiles based on users' historical job applications and behaviors. The system uses feature selection to analyze text information from applied jobs and integrates Support Vector Machine (SVM) for classification, ensuring personalized and relevant job recommendations	This research focuses on the domain of job recommendation systems within the broader field of machine learning and data science. It aims to assist job seekers, particularly college graduates, by aligning their skills and qualifications with suitable job opportunities in the technology industry	The dataset for this project was sourced from Kaggle, comprising student data. Key attributes include academic percentages, scores in algorithms, programming concepts, and coding ratings. This data is crucial for analyzing and recommending appropriate job opportunities based on	The system's performance is evaluated by its ability to provide accurate and relevant job recommendations. It leverages historical data and user profiles to ensure high precision in matching job seekers with suitable job roles. Continuous testing and refinement are integral to maintaining its effectiveness in real-	A significant limitation of the system is its dependency on the completeness and accuracy of user profiles. Inadequate or outdated information can impair the recommendation quality. Additionally, the system's effectiveness may be constrained by the diversity and scope of the dataset used for training and evaluation	The proposed job recommendation system effectively matches job seekers with suitable opportunities based on their skills and qualification. It highlights the potential of machine learning algorithms, particularly SVM, in enhancing the job search process. Future work aims to expand the dataset and refine the

			individual skill sets	world applications		recommendation algorithms for improved accuracy and relevance

[9]	The paper utilizes collaborative filtering techniques, specifically matrix factorization and k-nearest neighbors (KNN), to develop a scalable job recommender system. These methods are applied to user profiles, job descriptions, and behavioral data, integrating content-based and collaborative filtering to enhance recommendation accuracy and personalization.	The study is focused on the online job recommendation domain. It aims to match job seekers with suitable job postings by analyzing their resumes and behavior on job portals. This approach is applicable in various sectors like e-business, social media, and e-learning, where personalized recommendations are crucial	The research employs three datasets: (1) Random Dataset with 3,494 entries, manually annotated; (2) Feedback Dataset with 6,650 entries, based on user feedback; and (3) Candidates Dataset with 15,625 entries, containing job applications matching user profiles. An aggregated dataset of 26,669 entries is also used for comprehensive evaluation.	The system's performance is evaluated using metrics such as accuracy and precision. The hybrid approach combining collaborative and content-based filtering demonstrates satisfactory results across different datasets, outperforming traditional methods in terms of recommendation accuracy and efficiency.	The paper identifies several limitations: scalability issues, the cold-start problem, and sparsity in traditional collaborative filtering methods. Additionally, there are challenges with vocabulary control, tag ambiguity, and privacy concerns, which need to be addressed for more effective and secure recommendations.	The study concludes that hybrid recommender systems, which integrate multiple filtering techniques, offer improved performance in job recommendations. It highlights the need for better evaluation measures, scalability solutions, and enhanced privacy and security frameworks to address existing system gaps and provide more accurate and personalized job recommendations

[10]	The paper proposes a job recommendation system leveraging Deep Reinforcement Learning (DRL). It involves data collection from job seekers' profiles and job descriptions, preprocessing these data, feature extraction using neural networks, and applying DRL to optimize the recommendation process. The DRL model continuously learns and improves recommendations based on user interactions and feedback	The domain of this research is online job recommendation systems. It focuses on improving the accuracy and efficiency of matching job seekers with suitable job opportunities by using advanced machine learning techniques, specifically deep reinforcement learning, within the context of job portals and employment websites.	The paper utilizes various data sources including job seeker profiles, job descriptions, and historical interaction data. This data is preprocessed to remove noise and extract relevant features. The specific datasets include both structured and unstructured data, enabling the model to capture a comprehensive understanding of user preferences and job requirements.	The system's performance is evaluated using metrics like precision, recall, and mean average precision (MAP). The DRL-based recommendation system demonstrates superior performance compared to traditional methods, showing improvements in recommendation accuracy and user satisfaction through continuous learning and adaptation to user interactions.	The paper acknowledges limitations such as the need for large-scale datasets and significant computational resources. Additionally, there are challenges in model interpretability and handling data sparsity. Addressing these limitations requires further research and development to enhance the system's scalability and practical application.	The study concludes that using deep reinforcement learning significantly improves the accuracy and personalization of job recommendations. The proposed system effectively matches job seekers with relevant job postings, enhancing the overall job search experience. Future work aims to address existing challenges and further optimize the recommendation process.

[11]	The paper is based on the Knowledge Discovery from Data (KDD) model, involving data acquisition, anonymization, annotation, and feature extraction. Employed various multi-class classification algorithms to classify job types (real job, identity theft, corporate identity theft, multi-level marketing).	Detection and classification of fraudulent job advertisements in the online job market.	Dataset: Employment Scam Aegean Dataset (EMSCAD) consisting of 17,880 job vacancies collected between 2012 and 2014, annotated for fraud detection.	Achieving the best performance with a Gradient Boosting classifier that combined empirical rule-set based features, parts-of-speech tags, and bag-of-words vectors, with an F1-score of 0.88.	Existing methods for identifying fraudulent employment have limits in scalability, interpretability, and transparency. Common lexical properties may not be enough to capture the contextual semantics of employment adverts. Validation with a publicly accessible dataset may have constraints in terms of representativeness and label quality.	Developed a validated machine learning system for identifying categories of fraudulent job advertisements, highlighting the need for ongoing research and updated datasets.

[12]	The paper conducts data preprocessing, comparative analysis of machine learning classification techniques such as Bernoulli's Naïve Bayes, Multinomial Naïve Bayes, Random Forest, Linear SVM, and LSVM with elastic penalty.	Text classification within the domain of job titles and descriptions, aiming to improve the accuracy of candidate selection processes.	The dataset consists of job titles and descriptions sourced from Kaggle, with a total of 55,000 samples, including various job categories such as Administrative Assistant and Customer Service Representative.	Evaluation of the accuracy of the techniques, with Linear SVM achieving the best accuracy of 96.25% on 55,000 samples. Naïve Bayes classifiers demonstrated lower accuracy, indicating their limitations in this context.	Limitations of the paper are: Naïve Bayes classifiers exhibited poor classification performance, which may hinder their applicability in real-world scenarios. The computational requirements for training classifiers for each query document can be excessive, impacting efficiency.	The primary outcome measured in this study is the accuracy of different machine learning classification techniques in classifying job titles based on job descriptions.
[13]	The proposed system employs machine learning algorithms, including web crawling for data extraction, decision tree for candidate selection, and K-means clustering for job demand analysis involving data extraction, email notifications to job seekers, and clustering of job types.	Online job portals, focusing on optimizing job searching and hiring processes.	The data set consists of job circulars extracted from various company websites, including job requirements, experience needed, and salary offerings.	The web crawler demonstrated varying accuracy rates, achieving up to 100% accuracy in later tests, with response times improving significantly across test cases.	The paper has a limitation of initial tests showing the crawler's inability to gather data, indicating potential issues in the early stages of implementation.	The system successfully clusters similar job postings, sends targeted email notifications to job seekers, and enhances the efficiency of job searching and hiring processes.

[14]	This research conducts Data Processing: Transform unstructured data into a structured format, followed by textual processing and clustering using K-means. It matches job clusters with job seeker behavior attributes to provide personalized job recommendations.	Job recommendation systems utilizing natural language processing and machine learning techniques.	Job offers and job seeker interactions, including ratings, likes, and reviews, collected from job search websites.	The model aims to improve the matching of job offers to candidates based on their past interactions and preferences, though specific performance metrics are not detailed in the summary.	Potential limitations include the quality of scraped data, the effectiveness of clustering algorithms, and the need for continuous model training and evaluation.	The suggested methodology successfully groups job offers based on common characteristics and matches them with job seekers, improving the job search experience by making appropriate suggestions. Future work will concentrate on fine-tuning the model using sophisticated approaches such as Word2vec and assessing its performance.
[15]	This job recommendation system uses text pre-processing, TF-IDF vectorization, and cosine similarity to match resumes with job descriptions based on skill sets. It also suggests additional skills for career improvement.	Recommending IT jobs to college graduates and engineers.	The system utilizes datasets of resumes, job descriptions, and required skills for various IT roles.	The system ranks job recommendations based on similarity scores and offers skill-based improvement suggestions.	The system's accuracy depends on data quality, relies solely on text similarity, and is currently limited to the IT sector.	This system provides a personalized ranked list of job recommendations and career guidance for IT professionals based on their skill set.

[16]	The system employs text parsing, stop word removal, lemmatization, TF-IDF, cosine similarity, and K-Nearest Neighbors (KNN) to match resumes with job descriptions effectively.	Resume-based job recommendation systems.	Utilizes PDF resumes and job descriptions for training and evaluation.	The system significantly improves resume-to-job matching by leveraging NLP and deep learning, providing accurate and relevant job recommendations.	Limited by the need for manual input and potential inaccuracies in text extraction. Performance can vary based on resume format and job description quality.	Developed a robust system using advanced NLP techniques and deep learning, enhancing resume matching with job descriptions and offering actionable feedback for job seekers.
[17]	Combines machine learning and NLP techniques, including text preprocessing, feature extraction, and classification algorithms, to screen and rank candidates' resumes based on job descriptions.	Resume screening and candidate ranking in recruitment.	Employs a dataset of resumes and job descriptions for model training and evaluation.	Achieved high accuracy in candidate ranking and resume screening, improving the relevance of matches between resumes and job requirements.	Limited by the variability in resume formats and the need for a large annotated dataset to train the model effectively	Developed a hybrid model that enhances resume screening and candidate ranking by integrating machine learning with NLP, offering improved efficiency in the hiring process.

[18]	Employs deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for advanced resume parsing and job matching, aiming for higher precision and contextual understanding of resumes.	Focuses on resume parsing and job matching through deep learning, aiming to improve accuracy and relevance in the recruitment process.	Utilizes a dataset comprising diverse resumes and job descriptions, processed to train deep learning models and achieve better parsing and matching performance.	Demonstrated superior accuracy in parsing resumes and matching them with job descriptions compared to traditional methods, resulting in better alignment with job requirements and improved recruitment efficiency.	Faced challenges with handling diverse resume formats and the high computational resources required for training deep learning models, which can be a constraint for deployment.	Developed an advanced resume parsing system using deep learning techniques, offering enhanced job matching capabilities and improving the effectiveness of resume processing in recruitment
[19]	Applies Convolutional Neural Networks (CNNs) and transfer learning techniques for automated resume screening and ranking, focusing on extracting features and improving job candidate evaluations.	Centers on automated resume screening and ranking using advanced deep learning methods to streamline the recruitment process and enhance candidate selection accuracy.	Uses a diverse dataset of resumes and job descriptions, leveraging transfer learning to fine-tune models on specific resume and job matching tasks for improved performance.	Achieved notable improvements in resume screening and candidate ranking, demonstrating high accuracy and efficiency in processing and evaluating resumes compared to traditional methods.	Challenges include the need for significant computational resources and potential limitations in handling very diverse or unstructured resume formats effectively.	Created an effective automated resume screening and ranking system using CNNs and transfer learning, leading to better recruitment outcomes through improved resume evaluation and candidate matching.

[20]	Employs Convolutional Neural Networks (CNNs) for resume screening and ranking, integrating techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity to enhance feature extraction and candidate evaluation.	Focuses on resume screening and ranking within the recruitment domain, using deep learning methods to automate and improve the efficiency of candidate selection and evaluation processes	Utilizes a dataset of resumes and job descriptions, applying CNNs to extract features and rank resumes based on their relevance to job descriptions and required skills.	Demonstrates high accuracy and efficiency in resume screening and ranking, outperforming traditional methods by effectively handling and analyzing complex resume data through deep learning.	Potential limitations include the reliance on extensive training data and computational resources, as well as challenges in generalizing across varied resume formats and job descriptions.	Developed a robust resume screening and ranking system using CNNs, significantly improving the precision and speed of candidate selection processes through advanced deep learning techniques.

Categorization of Selected Domain and your perception about its importance:

The problem of categorizing machine learning job role listings on Glassdoor is mostly in the domain of prediction. It is specifically concerned with the prediction and classification of work positions based on features and patterns extracted from job listings. This categorisation can aid in anticipating candidates' suitability for specific work categories, as well as trends in job needs.

Prediction Domain:

- Prediction covers a wide range of applications in which the primary purpose is to forecast or estimate future outcomes using current and previous data. It employs strategies for analyzing patterns, trends, and relationships in data in order to create informed predictions.
- Predictions offer useful insights that can help influence strategic decisions. For example, in the context of job postings, forecasting future demand for specific talents might assist educational institutions in tailoring their curriculum accordingly.
- In job role classification, this involves providing job seekers with more relevant job listings, hence improving their job search experience.
- Enhances job matching by providing personalized job recommendations to candidates.
- Offers insights into job market dynamics and trends.

Description:

- The goal is to predict job roles and categorize listings based on the data extracted from Glassdoor.

Relevant Techniques:

- Support Vector Machine (SVM): Used for classification of job roles.
- Random Forest: Used for classification and regression tasks to identify patterns in job listings.
- Clustering: Used to group similar job listings together based on their features.

Applications:

- Predicting the type of job role based on the description and requirements.
- Classifying job listings into different categories (e.g., Data Scientist, Machine Learning Engineer, AI Researcher).
- Analyzing trends in job postings over time to predict future demand for specific roles.

Specific Importance to Machine Learning Job Role Listings:

1. **Skill Gap Analysis:**
    - Predicting which skills are in high demand can help job seekers and professionals upskill accordingly, reducing the skill gap in the market.
2. **Recruitment Strategy:**
    - Companies can refine their recruitment strategies by understanding which job roles are becoming more prevalent and what qualifications are most sought after.
3. **Enhanced Job Matching:**
    - Improving the accuracy of job role classification leads to better matching of candidates with job openings, increasing job satisfaction and retention rates.

Limitations/Gaps Identified in the Research Papers with Examples

1. **Sample Size**:
    - Small sample sizes can lead to overfitting and reduce the statistical power of the study. This limits the ability to draw meaningful conclusions and affects the model's performance on larger, more diverse datasets.
    - **Example**: The study on job position classification was limited by a sample size of 955 instances, which may not be sufficient to capture the full variability of job positions and their characteristics.
2. **Source Limitation**:
    - Reliance on data from a single source, such as a specific job board or online platform, may introduce source-specific biases. This limits the representativeness of the findings and reduces the applicability of the results to other sources or contexts.
    - **Example**: The content analysis of online job advertisements relied solely on data from Indeed.com, potentially missing relevant job postings from other sources, thereby limiting the comprehensiveness of the analysis.
3. **Data Bias and Dependency**:
    - Historical data and publicly available profiles often reflect past trends and may carry inherent biases. This can skew the model's predictions and reduce its accuracy in predicting future job market trends or behaviors.
    - **Example**: The job recommendation study used publicly available profiles that may reflect outdated skills and job titles, leading to biased recommendations that do not accurately represent current job market demands.
4. **Fairness and Generalizability**:
    - Algorithmic fairness is a critical issue, as models may inadvertently favor certain groups over others. Additionally, models trained on specific datasets may not generalize well to different datasets, reducing their applicability and reliability across various contexts.
    - **Example**: The job recommender systems study identified fairness issues where the algorithms performed differently across various demographic groups, and the models did not generalize well to datasets from different regions or industries.
5. **Complexity and Subjectivity**:

- Quantifying complex and subjective phenomena, such as organizational culture or job fit, poses significant challenges. These aspects are difficult to measure accurately, leading to potential inconsistencies and reduced validity of the findings.
- **Example**: The research on organizational culture faced difficulties in quantifying the inherently subjective and complex nature of organizational culture, leading to potential inconsistencies in the findings.

6. **Imbalanced Datasets**:
    - Handling imbalanced datasets, where certain classes are underrepresented, can lead to biased models that perform poorly on minority classes. This impacts the overall accuracy and fairness of the model's predictions.
    - **Example**: The classification of fake job postings likely dealt with an imbalanced dataset where legitimate job postings vastly outnumbered fake ones, resulting in a model that could struggle to accurately identify the minority class of fake postings.
7. **Dynamic and Temporal Nature**:
    - The dynamic and temporal nature of data, such as job openings that change frequently, requires models to adapt quickly. Static models may fail to capture these changes, leading to outdated or inaccurate predictions.
    - **Example**: Hybrid recommendation systems had to manage the dynamic nature of job openings, which change frequently. A static model might fail to adapt to these changes, resulting in outdated job recommendations.
8. **Cold-Start Problem and Sparsity**:
    - The cold-start problem occurs when there is insufficient data to make accurate recommendations for new users or items. Data sparsity, where there are few interactions or data points, further complicates the effectiveness of recommendation systems.
    - **Example**: Collaborative filtering techniques in job recommendation systems faced the cold-start problem when recommending jobs to new users with no prior interaction history, and sparsity issues where limited data points made it difficult to make accurate recommendations.
9. **Privacy and Security Concerns**:
    - Ensuring the privacy and security of sensitive personal information is crucial. Data breaches or mishandling of personal data can lead to significant ethical and legal issues, affecting the trustworthiness of the research.
    - **Example**: Studies on job recommender systems had to ensure the privacy and security of users' personal information, such as job application histories and preferences, which, if mishandled, could lead to ethical and legal challenges.
10. **Feature Extraction and Selection**:
    - Effective feature extraction and selection are critical for model performance. Poorly chosen features can lead to suboptimal models, while complex feature extraction processes can increase computational costs and reduce model interpretability.
    - **Example**: The research on job classification and recommendation often struggled with feature extraction and selection, where the complexity of identifying relevant features from job descriptions and user profiles could impact the performance and interpretability of the models.

<u>How the limitation from researchers needs to be resolved</u>:

Mitigating the limitations of prior studies is imperative for enhancing machine learning models' performance and applicability in the context of analyzing job role listings on Glassdoor.

1.Small Dataset Size

Limitations:

The lack of data in many studies is a serious issue, which results in overfitting and consequently, low generalizability.

Solutions:

1. **Data Augmentation**: Increase the size of the dataset by creating synthetic data through various methods.
2. **Web scraping**: To create a more varied and comprehensive dataset, keep pulling job listings from Glassdoor.
3. **Data Collaboration**: Work together to exchange and combine data resources with other institutions or researchers..
4. **Transfer Learning**: Take advantage of pre-trained models that are related to the task and fine-tune them on the job listings dataset to benefit from large datasets from similar domains.

2.Scalability Issues

Limitations:

Models seem to have a problem when it comes to dealing with large data sizes, which in turn causes inefficiencies to arise.

Solutions:

1. **Distributed Computing**: Using frameworks such as Apache Spark and Hadoop to implement a distributed data processing approach would be a great alternative for large data processing.
2. **Cloud Computing**: Get cloud platforms to support as many resources as needed in order to scale computing resources easily. For example AWS or Google Cloud.
3. **Efficient Algorithms**: Choose algorithms that are specifically designed to operate on large datasets efficiently, for instance, XGBoost for classification tasks.

3.Bias and Fairness

Limitations:

The AI's heavy reliance on unbalanced data may produce racy or wrongful results.

Solutions:

1. **Bias Detection and Mitigation**: Techniques, for instance, using resampling, reweighting, or adversarial debiasing, to detect and eliminate bias in the dataset should be applied.
2. **Fairness Metrics**: Evaluate the model by using fairness metrics and standardize the discipline of different pairs (e.g. demographic parity, equalized odds).
3. **Diverse Data Collection**: Diversity in the data might be achieved by including various job roles and industries in the dataset which may subsequently remove bias effects.

4.Feature Extraction

Limitation:

A factor influencing the model's accuracy is the difficulty of extracting features from job descriptions.

Solutions:

1. **Expert NLP Methods**: Use an extensive array of natural language processing (NLP) tools, such as Word2Vec, BERT, and GPT, among others, to generate rich feature representations of text.
2. **Domain-Specific Ontologies**: Create or utilize existing ontologies and taxonomies specific to certain job roles to improve the feature extraction process.
3. **Text Preprocessing**: Conduct extensive text preprocessing, encompassing NER (named entity recognition), stemming and lemmatization, to enhance feature quality.

5.Model Interpretability

Limitation:

The primary reason we are unable to comprehend neural networks at all is that they are non-linear deep networks that are obviously difficult to explain.

Solutions:

1. **Resources to Help with Interpretability**: Use tools to illustrate the explanations of the model predictions, such as SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations).
2. **Simple Models**: In a given case where it is feasible, it is better to select simple models such as decision trees, or methods like linear regression, which are not complex in nature and therefore more interpretable.
3. **Model Explainability**: Enhancing the transparency of the model by using explainability frameworks will be the solution to the problem, which will, in turn, help the stakeholders to comprehend how the model makes its decisions.

6.Handling Unbalanced Data

Limitations:

Unbalanced data from job listings, such as when a role's number is greater than others.

Solutions:

1. **Techniques for Resampling**: To balance the dataset, apply undersampling or oversampling techniques (such as SMOTE).
2. **Class Weights**: Adjust the learning algorithm so that the underrepresented groups are given varying class weights to indicate their relative importance.
3. **Anomaly Detection**: To identify and handle uncommon job roles as the anomaly, apply anomaly detection techniques.

7.Evaluation Metrics and Validation:

Limitations:

Inadequate validation methods and evaluation metrics can cause an algorithm's performance to be estimated incorrectly.

Solutions:

1. **Robust Validation**: To make sure the model is generalizing well to new, unseen data and not just memorizing the training set, cross-validation should be employed.
2. **Comprehensive Metrics**: The overall assessment of the performance can be obtained by means of the multi-metrics evaluation, such as, accuracy, precision, recall, and F1 score for classification; silhouette score and Davies-Bouldin index for clustering.
3. **Real-World Testing**: The experimental tests can be performed in the real-world settings with the live data in order to validate the model performance in practice.

8. Data Privacy and Security

Limitations:

Keeping data private and secure is one of the biggest challenges especially in connection with sensitive job-related data.

Solutions:

1. **Data Anonymization**: The job offers must not contain any personal information in order to protect privacy.
2. **Secure Data Storage**: Use strong security controls for data storage and processing, such as encryption and access control, to back up your data.
3. **Compliance**: Ensure compliance with data protection laws (e.g., GDPR, CCPA) when processing job listing data to make your company trustworthy.

We can improve the robustness, scalability, and fairness of your machine learning models and obtain more accurate and trustworthy insights from Glassdoor job listings by leveraging the gaps that have been identified and the solutions that have been proposed.

Concluding Remarks

Significance of the Selected Domain:

The domain of Prediction, specifically in the context of categorizing machine learning job role listings on Glassdoor, is highly significant for several reasons:

1. **Enhanced Job Matching:**
    - **Candidate Suitability:** By accurately predicting and classifying job roles, candidates can be better matched with positions that suit their skills and experiences. This leads to a more efficient job search process and increases the chances of job satisfaction and retention.
    - **Employer Efficiency:** Employers can streamline their hiring process by identifying and prioritizing candidates who are best suited for specific roles, thus reducing time-to-hire and improving overall recruitment efficiency.
2. **Trend Analysis:**
    - **Market Insights:** Analyzing trends in job postings over time helps in understanding the evolving demands of the job market. This information is invaluable for both job seekers and educational institutions to align their skills and curriculum with market needs.
    - **Future Demand:** Predicting future demand for specific roles enables proactive career planning for individuals and strategic workforce planning for organizations.
3. **Skill Development:**
    - **Educational Guidance:** Educational institutions can use these insights to update their programs and courses, ensuring they are aligned with industry requirements. This prepares students better for the job market, enhancing their employability.
    - **Professional Growth:** Professionals can identify emerging skills and roles, allowing them to upskill and stay competitive in their careers.

Final Thoughts:

The task of categorizing machine learning job role listings using SVM, Random Forest, and Clustering is a compelling application of predictive analytics. It demonstrates the power of machine learning in transforming raw data into actionable insights. By addressing the limitations and leveraging the strengths of these techniques, we can significantly enhance the efficiency and effectiveness of the job market for both candidates and employers. This project not only contributes to the academic and professional growth of students but also provides practical solutions to real-world challenges in the employment sector.

ML (1).docx

ORIGINALITY REPORT

7%	3%	3%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Amir Shachar. "Introduction to Algogens", Open Science Framework, 2024 Publication	1%
2	papers.ssrn.com Internet Source	<1%
3	Submitted to Sheffield Hallam University Student Paper	<1%
4	www.irjmets.com Internet Source	<1%
5	Marcel Naudé, Kolawole John Adebayo, Rohan Nanda. "A machine learning approach to detecting fraudulent job types", AI & SOCIETY, 2022 Publication	<1%
6	Submitted to Reykjavík University Student Paper	<1%
7	insights2techinfo.com Internet Source	<1%
8	Submitted to ESCP-EAP Student Paper	<1%

9	Submitted to University of the West Indies Student Paper	<1 %
10	J. Himabindu Priyanka, Nikhat Parveen. "DeepSkillNER: An automatic screening and ranking of resumes using hybrid deep learning and enhanced spectral clustering approach", Multimedia Tools and Applications, 2023 Publication	<1 %
11	journals.nawroz.edu.krd Internet Source	<1 %
12	Submitted to College of Estate Management Student Paper	<1 %
13	Submitted to University of Exeter Student Paper	<1 %
14	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
15	www.mdpi.com Internet Source	<1 %
16	Submitted to Jose Rizal University Student Paper	<1 %
17	Submitted to Kaplan College Student Paper	<1 %
18	Submitted to University of Sunderland	

Student Paper

<1%

19	Ginel Dorleon, Nathalie Bricon-Souf, Imen Megdiche, Olivier Teste. "Absolute Redundancy Analysis Based on Features Selection", 2021 4th International Conference on Data Science and Information Technology, 2021 Publication	<1%
20	ijsr.net Internet Source	<1%
21	practicaldatascience.co.uk Internet Source	<1%
22	www.fastercapital.com Internet Source	<1%
23	R Maruthaveni, B Dhivya, M Hariharan, R Siva. "FIREnet 2.0: Advanced Neural Framework for Smart Fire Detection & Localization", 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), 2023 Publication	<1%
24	D. Mhamdi, R. Moulouki, M.Y. El Ghoumari, M. Azzouazi, L. Moussaid. "Job Recommendation based on Job Profile Clustering and Job Seeker Behavior", Procedia Computer Science, 2020	<1%

Publication

25	www.frontiersin.org Internet Source	<1%
26	Samta Jain Goyal, Rajeev Goyal, Vinay Kumar Singh, Rajesh Arunachalam, Kuldeep Narayan Tripathi. "Privacy-preserving cross-domain recommendation using hybrid federated transfer learning", Multimedia Tools and Applications, 2024 Publication	<1%
27	fastercapital.com Internet Source	<1%
28	id.123dok.com Internet Source	<1%
29	Ton Duc Thang University Publication	<1%
30	Vedant Das Swain, Koustuv Saha, Manikanta D. Reddy, Hemang Rajvanshy, Gregory D. Abowd, Munmun De Choudhury. "Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

ML (1).docx

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22