iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: VIII Month of publication: August 2024

DOI: https://doi.org/10.22214/ijraset.2024.63857

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VIII Aug 2024- Available at www.ijraset.com

Detecting Fraudulent Job Postings: An Analysis of Machine Learning Approaches

Rozia Razak[1], Dr. Gurinder Kaur Sodhi[2]
[1]M. Tech Scholar, Department of Electronics and Communication Engineering, Desh Bhagat University, Mandi Gobindgarh Punjab, India
[2]Professor, Department of ECE, Desh Bhagat University, Punjab, India

***Abstract:** The rise of online job platforms has greatly simplified the global job search process for millions of individuals. However, this convenience has also led to a concerning issue - an increase in fraudulent job postings. Job seekers are becoming more susceptible to scams, identity theft, and financial fraud as malicious entities post deceptive job advertisements. This paper conducts a thorough analysis of machine learning approaches aimed at detecting fraudulent job postings, with the main goal of improving the legitimacy and safety of online job markets. The study encompasses data collection, preprocessing, feature engineering, model selection, and evaluation. To tackle these challenges, we explore the effectiveness of various machine learning models, including Decision Trees, Random Forests, and Support Vector Machines (SVM). Additionally, we investigate data preprocessing techniques such as label encoding and standardization to prepare the data for modeling. The results suggest that while Decision Trees and Random Forests show promising performance, dealing with the imbalanced nature of the dataset requires the application of oversampling techniques for enhanced accuracy in identifying fraudulent postings. The paper provides a comprehensive evaluation of model performance, utilizing metrics like accuracy, F1-score, precision, and recall. Furthermore, it emphasizes the significance of strategies to handle class imbalance in real-world applications.*

***Keywords:** Fraud Jobs, EDA, Random forest, Machine Learning*

I. INTRODUCTION

Over the past few decades, the digital revolution has brought about profound changes in different aspects of society, notably in how people search for and obtain employment. The emergence of online job markets signifies a substantial shift in the landscape of job searching. This section delves into the evolution and widespread adoption of online job markets, elucidating the factors that have played a pivotal role in their increased prominence..

Figure 1 Fraudulent job fishing

A. *Technological Advancements and Connectivity*

The internet's advent and the widespread availability of high-speed connectivity have fundamentally altered the conventional job-seeking process. The accessibility of the internet, combined with the prevalence of digital devices, has granted job seekers the ability to explore employment opportunities from virtually any location worldwide. This technological advancement has erased geographical barriers and brought about a revolutionary shift in how job seekers establish connections with potential employers.

B. *Convenience and Accessibility*

Online job markets present an unprecedented level of convenience for job seekers. These platforms feature user-friendly interfaces that enable individuals to peruse job listings, submit applications, and engage with employers without the limitations of physical proximity. Job seekers can explore a myriad of job postings, make comparisons between opportunities, and customize their searches to align with their career objectives, all from the comfort of their homes or mobile devices..

C. Diverse Job Opportunities

Online job markets are not confined to a single industry or sector. They encompass a wide spectrum of job opportunities, ranging from entry-level positions to executive roles and spanning various domains, including technology, healthcare, finance, and creative fields. This diversity of job listings caters to the diverse skills and aspirations of job seekers.

In the digital age, the evolution of online job markets has revolutionized the way individuals explore, apply for, and secure employment opportunities. Online job platforms have become essential tools for job seekers, offering a wide range of job listings across various industries and geographies. However, within the vast array of legitimate job postings, there exists a concerning presence of fraudulent job advertisements. The proliferation of these deceptive listings has raised doubts about the credibility of online job markets, posing a threat to the financial and personal security of job seekers. This paper is motivated by the pressing need to develop and implement a robust fraud detection system within online job markets. Such a system, leveraging the capabilities of machine learning, holds the potential to effectively differentiate between genuine and fraudulent job postings, ensuring the safety of job seekers and maintaining the integrity of online job platforms.

The paper conducts a thorough exploration of the intricate landscape of fraudulent job postings, delving into the methodologies, techniques, and machine learning algorithms employed to identify and mitigate the risks associated with deceptive job listings. Additionally, it addresses the challenge of class imbalance within the dataset, a common issue in fraud detection, and discusses effective strategies for managing this imbalance.Going beyond technical considerations, the research extends into the practical realm by examining the deployment of the fraud detection system in a real-time production environment. It explores the development of a user-friendly interface and an alerting system, facilitating swift assessments of job postings to promptly shield job seekers from potential fraud.

II. LITERATURE REVIEW

Vidros et al. (2016), who diferentiated between two groups of fraudulent jobs. The frst group comprises advertisements for non-existing jobs which aim to harvest personal information such as names, phone numbers, and e-mail addresses. Such information may then be sold to third parties or used as targets for spam emails and spam calls. The second group of fraudulent jobs consists of attempts to social engineer either highly sensitive information out of the job seeker, such as social security numbers or passports, or lure the job seeker into depositing sums of money.

Reynolds (2021) wrote about nine diferent types of job search scams, which difer in the type of actions it demands from the job seeker. JobHunt, a career advice website, gives an example of ‘corporate identity theft’, which are fraudulent jobs that claim to be from a real employer (Joyce 2021

III. OBJECTIVES

1) Develop a machine learning-based system to accurately detect fraudulent job postings within online job markets.
2) Prioritize job seeker security by promptly alerting them to potential risks associated with fraudulent listings.
3) Restore trust in online job platforms by reducing the prevalence of deceptive job postings.
4) Address class imbalance in the dataset to improve model accuracy.
5) Implement a feedback loop mechanism for continuous system improvement over time.

IV. METHODOLOGY

In the initial phase of the system implementation, data was gathered from various online job platforms and stored in a centralized database or data repository. This data encompassed a wide range of details related to job titles, descriptions, locations, company profiles, and other pertinent information. Rigorous data preprocessing was undertaken to ensure the quality and consistency of the dataset. Tasks such as data cleaning, handling missing values, and outlier detection were executed meticulously to prepare the data for subsequent analysis and modeling, ensuring its accuracy and structured format.

The pivotal step of feature engineering followed, during which the dataset underwent transformations to extract meaningful information usable by the machine learning models. This involved techniques such as text processing to extract keywords and phrases from job descriptions. Additionally, categorical variables like employment type and required education were encoded, and numerical features were generated based on the available data. Feature engineering aimed to capture the most relevant aspects of job postings indicative of fraudulent or genuine listings.The core of the system comprised machine learning models selected for classification purposes. Various algorithms, including Decision Trees, Random Forests, and Support Vector Machines, were employed.

These models were carefully trained on the preprocessed and engineered dataset, enabling them to learn intricate patterns and relationships between features that could effectively distinguish between genuine and fraudulent job postings. Hyperparameter tuning was conducted to optimize the models' performance, ensuring accurate predictions. Once the models were fine-tuned and validated, they were deployed into a production environment. The deployment infrastructure included elements such as web servers or cloud-based services facilitating user interaction. Users engaged with the system through a user interface (UI), submitting job postings for analysis. If a job posting was classified as potentially fraudulent, an alerting system was triggered, facilitating notifications to users or administrators for further action. Operating in real-time, the system provided swift and accurate assessments of job postings, enhancing user safety and building trust in online job markets. The system was thoughtfully designed to include a feedback loop where user input and performance data were consistently collected. This data played a crucial role in the ongoing enhancement of the system, being leveraged for the periodic retraining of models, ensuring continuous improvement in accuracy over time. Additionally, a dedicated database was maintained for logging and reporting, enabling systematic monitoring and auditing of the system's performance and outcomes..

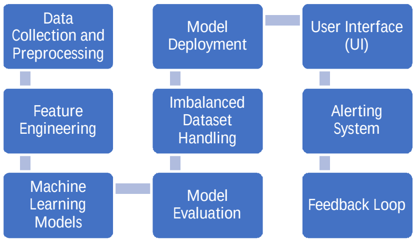

Figure 2 Flow diagram of the system

A. *Data Collection and Preprocessing*

Data collection and preprocessing serve as the initial building blocks in constructing an effective fraud detection system. This critical phase involves gathering and refining the dataset, laying the groundwork for subsequent model development and analysis.In the realm of data acquisition, a comprehensive dataset is procured from various online job markets. This dataset must encompass a wide spectrum of job postings, ranging from legitimate to fraudulent, and span diverse industries and job types. By doing so, the dataset captures the inherent variability present in online job listings, enabling the system to develop a nuanced understanding of fraudulent patterns. Once the dataset is assembled, rigorous data cleaning and sanitization processes come into play. This entails the removal of duplicate entries to maintain data integrity. Additionally, the handling of missing values, particularly in attributes like location, department, and salary range, is addressed through imputation or removal, ensuring the dataset's quality remains uncompromised.

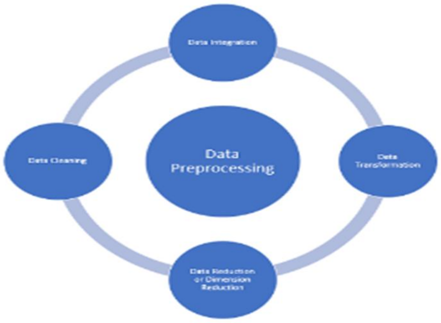

Figure 3. Data preprocessing

To prepare textual data for analysis, text preprocessing techniques are employed. This involves the transformation of job descriptions, requirements, and other text-based information into a format amenable to natural language processing (NLP). Techniques such as tokenization, stop-word removal, stemming, and vectorization are applied to facilitate meaningful feature extraction from textual data. A paramount concern in fraud detection is class imbalance, as fraudulent job postings often represent a minority class. To mitigate this issue, various resampling techniques are employed. Oversampling of the minority class and under sampling of the majority class are strategies employed to balance class representation and enhance model performance.

Feature engineering also plays a pivotal role in this phase. Relevant features, encompassing both numerical attributes like telecommuting and categorical attributes like employment type, required experience level, and education requirements, are extracted from the dataset. The preprocessed dataset is then partitioned into training and testing subsets, setting the stage for model training and evaluation. Standardization of numerical features ensures uniform scaling, preventing any particular attribute from exerting undue influence during modeling.

Further, the dataset is enriched with labels that indicate the authenticity of each job posting, distinguishing between fraudulent and legitimate listings. Exploratory Data Analysis (EDA) is performed to glean insights into the dataset's characteristics, unveil patterns, and visualize the distribution of fraudulent and legitimate job postings.Data validation serves as the final checkpoint in this phase, ensuring that the preprocessed dataset is ready for training and evaluating machine learning models. This meticulous preparation lays the foundation for subsequent stages of model development, fostering confidence in the dataset's quality and readiness for fraud detection analysis

B. Data Sources

The success of any fraud detection system hinges on the quality and diversity of the data sources used. In this section, we delve into the sources from which the dataset for fraudulent job posting detection is collected, highlighting the significance of a comprehensive and varied dataset.

The primary data source is drawn from multiple online job markets. These platforms serve as repositories for a plethora of job listings across industries, positions, and geographical locations. To ensure the dataset's representativeness, job postings are extracted from a range of online job markets, each contributing to the overall diversity of the data.

V. EXPERIMENTAL SETUP

A. Data Cleaning and Quality Assurance

Ensuring the integrity and reliability of the dataset is paramount in the development of a fraud detection system. This section delves into the critical processes of data cleaning and quality assurance, which are essential to maintain the dataset's accuracy and consistency.

1) Data Cleansing

Data cleansing involves a meticulous review of the dataset to identify and rectify anomalies, errors, and inconsistencies. Key steps in this process include:

a) *Duplicate Entry Removal:* Duplicate job postings, if present, are identified and eliminated to prevent skewed analysis due to redundancy.

b) *Handling Missing Values:* Missing values in attributes such as location, department, and salary range are addressed through appropriate methods. This may involve imputing missing values with sensible defaults or removing incomplete records.

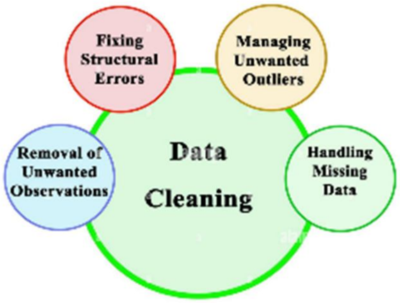

Figure 4 Data cleaning

2) Quality Assurance

Quality assurance measures are implemented to validate the dataset's quality and readiness for analysis:

a) *Data Validation:* Rigorous validation procedures are applied to confirm the accuracy and consistency of the dataset, ensuring that it adheres to predefined standards.

b) *Outlier Detection:* Outliers, if any, are identified and reviewed. Outliers can indicate potential errors or anomalies in the data that need further investigation.

3) Textual Data Preprocessing:

As textual data often plays a central role in fraud detection, preprocessing of text-based attributes is crucial:

a) Tokenization: Text is broken down into individual tokens or words for subsequent analysis.

b) Stop-Word Removal: Commonly used words (stop words) that carry little semantic meaning are removed to focus on essential content.

c) Stemming: Words are reduced to their root form (stem) to standardize language variations.

d) Vectorization: Text data is transformed into numerical vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings.

4) Handling Class Imbalance:

Addressing class imbalance between legitimate and fraudulent job postings is vital for model training:

a) Oversampling: The minority class (fraudulent job postings) may be oversampled to increase its representation in the dataset.

b) Undersampling: Conversely, undersampling the majority class (legitimate job postings) can balance class distribution.

c) Data Standardization: To ensure uniformity, numerical attributes are standardized:

d) Scaling: Numerical features are scaled to have a consistent range, preventing any single attribute from dominating model training.

5) Data Validation and EDA:

Data validation procedures are followed to confirm that the preprocessed dataset is suitable for analysis. Exploratory Data Analysis (EDA) is performed to gain insights into dataset characteristics, identify patterns, and visualize the distribution of fraudulent and legitimate job postings.

In summary, data cleaning and quality assurance are foundational steps in the development of a fraud detection system. These processes are indispensable for maintaining the dataset's accuracy, consistency, and readiness for subsequent stages of model development and analysis.

B. Handling Missing Values

Addressing missing values is a critical aspect of data preprocessing to ensure the integrity and completeness of the dataset. In this section, we delve into the strategies and considerations employed to handle missing values effectively.

1) Identification of Missing Values

The process begins with identifying missing values within the dataset. Missing values can occur in various attributes, such as location, department, salary range, and others. Thorough data examination reveals the extent of missing data in each attribute.

2) Strategies for Handling Missing Values

a) Imputation: For attributes where missing values can be reasonably estimated, imputation techniques are applied. This involves replacing missing values with statistically derived estimates, such as mean, median, or mode. Imputation ensures that the affected attributes remain part of the dataset and contribute to subsequent analyses.

b) Removal: In cases where missing values cannot be reliably imputed, removal of the corresponding records is considered. This approach is chosen when the missing values do not constitute a substantial portion of the dataset, and their absence does not significantly impact the analysis.

3) Imputation Techniques

The choice of imputation technique depends on the nature of the attribute and the dataset's specific characteristics:

a) Mean/Median Imputation: This approach is suitable for numerical attributes, where the missing values are replaced with the mean (average) or median of the available data. It maintains the central tendency of the distribution.

b) Mode Imputation: Mode imputation is employed for categorical attributes. Missing values are replaced with the most frequently occurring category, preserving the mode of the attribute.

C. Dataset Structure and Features

Understanding the structure and features of the dataset is essential for building a robust fraud detection system. This section provides an overview of the dataset's composition and the key features that will be leveraged in subsequent analyses.

1) Dataset Structure

The dataset is organized into rows and columns, with each row representing an individual job posting and each column representing a specific attribute or feature associated with that posting. This structured format facilitates systematic analysis and modeling. The dataset's structure encompasses the following components:

a) Rows: Each row corresponds to a single job posting entry, and the dataset contains a multitude of these entries, capturing a wide range of job listings.

b) Columns: The columns represent various attributes and characteristics associated with each job posting. These attributes encompass both numerical and categorical features, offering a comprehensive view of each listing.

2) Key Features

Several key features within the dataset play a pivotal role in the fraud detection process. These features are integral to the system's ability to differentiate between legitimate and fraudulent job postings.

Some of the essential features include:

a) Telecommuting: A binary attribute indicating whether the job allows telecommuting, which can be a potential indicator of legitimacy.

b) Company Logo: A binary attribute denoting whether the job posting includes a company logo, which may influence job seeker trust.

c) Questions: Another binary attribute signifying whether the job posting includes questions, potentially affecting job seeker engagement.

d) Employment Type: A categorical feature describing the type of employment offered, such as full-time, part-time, contract, etc.

e) Experience Level: This categorical feature indicates the required experience level for applicants, ranging from entry-level to executive.

f) Education Requirements: A categorical feature specifying the educational qualifications expected from applicants, including degrees and certifications.

g) Fraudulent Label: A binary label designating whether the job posting is fraudulent (1) or legitimate (0).

These features collectively provide a rich set of information for analysis and modeling. The binary attributes capture job posting characteristics, while the categorical features offer insights into job requirements and attributes that job seekers may consider when evaluating opportunities.

Understanding the dataset's structure and key features serves as the foundation for subsequent data analysis, model development, and the ultimate goal of effectively detecting fraudulent job postings within online job markets.

D. Data Visualization and Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a pivotal phase in the development of a fraud detection system, providing insights into the dataset's characteristics, revealing patterns, and aiding in feature selection. This section outlines the techniques and visualizations employed to gain a comprehensive understanding of the data.

1) Histograms and Distributions

Histograms are instrumental in visualizing the distributions of numerical attributes. By plotting histograms of features like job posting attributes, experience levels, and education requirements, we can discern the distribution shapes and identify potential outliers. Understanding these distributions is essential for feature engineering and model selection.

2) Bar Charts

Bar charts are utilized to visualize the distribution of categorical attributes such as employment type, required experience, and education requirements. These charts provide a clear representation of the frequency of different categories, aiding in the identification of dominant trends and imbalances within the dataset.

3) *Correlation Analysis*

Correlation matrices and heatmaps are employed to examine the relationships between attributes. This analysis helps uncover correlations between features, highlighting potential interactions that may influence the presence of fraudulent job postings. For instance, understanding whether telecommuting is correlated with job legitimacy can provide valuable insights.

4) *Box Plots*

Box plots are valuable for visualizing the spread and variability of numerical attributes. These plots reveal potential outliers and variations in features like salary range, which can be indicative of unusual job postings that warrant further investigation.

5) *Time Series Analysis (if applicable)*

For datasets with temporal components, time series plots can reveal trends and patterns in fraudulent activities over time. This analysis can uncover seasonality or temporal anomalies related to fraudulent job postings.

6) *Data Visualization Goals*

The overarching goal of data visualization and EDA is to uncover insights and patterns that inform subsequent stages of the fraud detection system's development. These insights guide feature selection, resampling strategies, and model design. Additionally, EDA helps in addressing data quality issues and ensuring that the dataset is well-prepared for model training and evaluation.

E. *Feature Engineering*

Feature engineering is a crucial phase in the development of an effective fraud detection system, as it involves the art of creating, transforming, and selecting attributes that enable machine learning models to distinguish between fraudulent and legitimate job postings accurately. In this project, several feature engineering techniques and strategies are applied to extract valuable insights from the dataset and improve model performance.

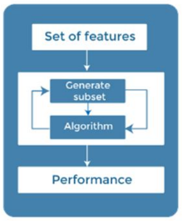

Figure 5 Feature selection

Temporal components, if present in the dataset, are also leveraged for feature engineering. Features related to posting timestamps, such as the day of the week, time of day, or temporal patterns in posting frequency, are generated. Temporal anomalies can provide insights into fraudulent activities that exhibit unusual posting patterns over time.

Finally, the feature engineering process is iterative, involving experimentation with different feature sets and transformations. Validation techniques such as cross-validation are employed to assess the impact of engineered features on model performance. The iterative refinement of features ensures that the most discriminative and relevant attributes are retained, ultimately enhancing the fraud detection system's effectiveness in distinguishing fraudulent from legitimate job postings.

F. *Feature Selection and Extraction*

In the quest to develop an effective fraud detection system, feature selection and extraction stand as critical steps in the data preprocessing pipeline. These processes are pivotal in refining the dataset, enhancing model performance, and reducing the computational complexity associated with high-dimensional data. In this project, a combination of feature selection and extraction techniques is employed to identify and retain the most informative attributes while mitigating issues such as multicollinearity and overfitting.One of the fundamental techniques in feature selection is the use of filter methods. These methods assess the relevance of each feature independently of any specific machine learning algorithm. Statistical tests, such as chi-squared tests for categorical features and correlation analysis for numerical attributes, are applied to quantify the relationships between features and the target variable—fraudulent job postings in this context. Features that exhibit low statistical significance or weak correlations with the target variable may be considered for removal.

This helps streamline the dataset, removing attributes that are unlikely to contribute significantly to the fraud detection task.Wrapper methods, on the other hand, involve the evaluation of feature subsets with respect to a specific machine learning algorithm. Recursive feature elimination (RFE) and forward selection are examples of such techniques. They iteratively assess subsets of features and gauge their impact on model performance. By selecting the most predictive feature set, wrapper methods optimize the model's ability to distinguish between fraudulent and legitimate job postings. This iterative approach is valuable in identifying the most relevant attributes while discarding those that offer little discriminatory power.

G. Text Processing for Job Descriptions

The comprehensive text processing techniques employed for job descriptions play a pivotal role in enhancing the fraud detection system's overall effectiveness. Job descriptions are a rich source of information that can provide valuable clues regarding the legitimacy of a job posting. By extracting meaningful insights from the textual content, the system becomes more adept at distinguishing between fraudulent and legitimate job offers.

VI. RESULTS AND DISCUSSION

The work begins by importing necessary libraries for handling numerical operations (numpy), data processing (pandas), visualization (seaborn and matplotlib), and operating system commands (os).

It sets the default figure size for plots to be 10 by 8 inches.The work reads a CSV file located at '/Users/User/Desktop/rozia/fake_job_postings.csv' into a DataFrame (a type of data structure in pandas for handling tabular data) named df.

Finally, it displays the first few rows (head) of the DataFrame (df) to provide a glimpse of the data:

Table 1 Dataset and parameters

	job_id	title	location	required_education	industry	function	fraudulent
0	1	Marketing Intern	US, NY, New York	NaN	NaN	Marketing	0
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	NaN	Marketing and Advertising	Customer Service	0
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	NaN	0
3	4	Account Executive - Washington DC	US, DC, Washington	Bachelor's Degree	Computer Software	Sales	0
4	5	Bill Review Manager	US, FL, Fort Worth	Bachelor's Degree	Hospital & Health Care	Health Care Provider	0

This dataset (df) has 17,880 rows and 18 columns. This Table provides information about our dataset, like the number of entries (rows), the types of data in each column, and how much memory it is using. It also shows us which columns have missing valuesThe table gives us basic statistics about the numerical columns in our data, such as the average (mean), spread (standard deviation), and the minimum and maximum values. It helps us understand the distribution of the numerical data in our dataset.

Table 2 Dataset categorised into percentage and various types of offers

	count	mean	std	min	25%	50%	75%	max
job_id	17880.0	8940.500000	5161.655742	1.0	4470.75	8940.5	13410.25	17880.0
telecommuting	17880.0	0.042897	0.202631	0.0	0.00	0.0	0.00	1.0
has_company_logo	17880.0	0.795302	0.403492	0.0	1.00	1.0	1.00	1.0
has_questions	17880.0	0.491723	0.499945	0.0	0.00	0.0	1.00	1.0
fraudulent	17880.0	0.048434	0.214688	0.0	0.00	0.0	0.00	1.0

This Table shows the number of missing values in each column of our dataset. For example, the "department" column has 11,547 missing values, while the "salary_range" column has 15,012 missing valuesThen we create a heatmap using seaborn (sns) to visualize the missing values in our dataset. The areas with a color indicate missing values. The y-axis labels (yticklabels) are turned off for better clarity in the visualization. The heatmap is a quick way to see the distribution of missing values across different columns

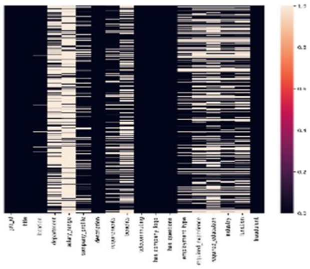

Figure 6 Heatmap

We check the length (number of entries) of the 'department' column in the dataset, which is 17,880and calculate and print the percentage of missing values for each feature (column) in the dataset. For example, the 'department' feature has 64.58% missing values We, then providea count of unique job titles in the 'title' column, showing how many times each title appears in the dataset.A count of unique locations in the 'location' column, showing how many job postings are associated with each location is also rovided. Then, a count of unique departments in the 'department' column, showing how many job postings are associated with each department.a count of fraudulent (1) and non-fraudulent (0) job postings in the 'fraudulent' column. There are 866 instances of fake job profiles.

We create a count plot using seaborn (sns) to visualize the distribution of job postings with and without telecommuting options. The plot is showing the count of each category on the x-axis.

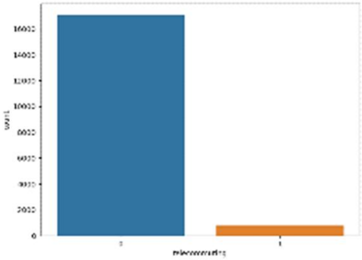

Figure 7 Telecommuting vs count graph

A plot is mapped to visualize the distribution of job postings with and without telecommuting options, taking into account whether the job posting is fraudulent or not. The plot displays the count of each category, with different colors representing fraudulent and non-fraudulent job postings.

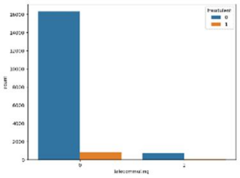

Figure 8 Detection of counts of fraudulent

From the above figure it is clear that most of the job profiles which does not have any telecommuting are not fake

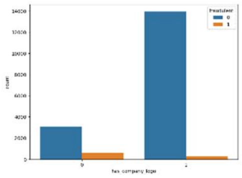

Figure 9 Company logo and fraudulent frequency

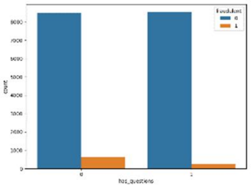

Figure 10 Questions and fraudulent frequency

we initially examined the 'employment_type' column to understand the distribution of different employment types, revealing insights such as 11,620 full-time positions and 1,524 contract positions. Subsequently, we identified and noted 3,471 missing values within the 'employment_type' column. To address this, we removed the corresponding rows with missing 'employment_type' values, resulting in a dataset of 14,409 rows and 18 columns. Additionally, we conducted a similar analysis for the 'required_experience' column, identifying and removing rows with missing values in this context. The 'required_education' column underwent a similar process, with the elimination of rows containing missing educational information. To streamline the dataset further, we compiled a list of unnecessary columns and successfully removed them from the dataset. These steps were undertaken to enhance the dataset's cleanliness and facilitate more effective analysis and modeling.

Table 3Features and requirements of the jobs

	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	fraudulent
3	0	1	0	Full-time	Mid-Senior level	Bachelor's Degree	0
4	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	0
6	0	1	1	Full-time	Mid-Senior level	Master's Degree	0
9	0	1	0	Part-time	Entry level	High School or equivalent	0
10	0	0	0	Full-time	Mid-Senior level	Bachelor's Degree	0

This displays the first few rows of the modified dataset after converting categorical values into numerical representations and resetting the index.

Table 4 Number of samples and the details present

	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	fraudulent
0	0	1	0	1	5	1	0
1	0	1	1	1	5	1	0
2	0	1	1	1	5	5	0
3	0	1	0	3	2	4	0
4	0	0	0	1	5	1	0

```
Fraudulent
0    7974
1     360
Name: count, dtype: int64
```

The dataset is an imbalanced dataset. So RandomOverSampling should be applied on the dataset

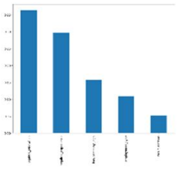

Figure 11 Classification tree results

A. Model Creation

The code calculated and prints the accuracy and F1 score of the Decision Tree classifier on the test data. The accuracy is approximately 95.44%, and the F1 score is about 0.17.

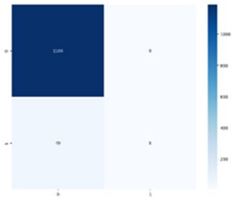

Figure 12 Confusion matrix

```
In [50]:
print(classification_report(y_test,yhat))
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      1196
           1       0.43      0.11      0.17        55
    accuracy                           0.95      1251
   macro avg       0.69      0.55      0.58      1251
weighted avg       0.94      0.95      0.94      1251
```

In [51]:
This code calculates and prints the accuracy and F1 score of the Random Forest classifier on the test data. The accuracy is approximately 95.76%, and the F1 score is about 0.18.>

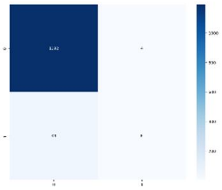

Figure 13 Confusiom matrix

This code calculates and prints the accuracy and F1 score of the SVM classifier on the test data. The accuracy is approximately 95.60%, and the F1 score is 0.0.

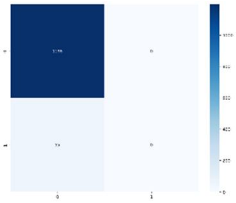

Figure14 SVM confusion matrix

The classification report for the Support Vector Machine (SVM) classifier on the test data indicates precision, recall, and F1-score for both classes (0 and 1), along with support and accuracy metrics. However, it raises an UndefinedMetricWarning because precision and F1-score are ill-defined and set to 0.0 in labels with no predicted samples.

B. *Interpretation of Model Outputs*

1) *Decision Tree Classifier*

- Accuracy: 95.44%
- F1 Score: 0.174
- Precision Score: 0.429
- Recall Score: 0.109
- True Positives: 6
- True Negatives: 1189
- False Positives: 10
- False Negatives: 46

2) *Random Forest Classifier*

- Accuracy: 95.76%
- F1 Score: 0.185
- Precision Score: 0.6
- Recall Score: 0.109
- True Positives: 6
- True Negatives: 1190
- False Positives: 6
- False Negatives: 49

3) *Support Vector Classifier (SVC)*

- Accuracy: 95.60%
- F1 Score: 0.0
- Precision Score: 0.0

- Recall Score: 0.0
- True Positives: 0
- True Negatives: 1196
- False Positives: 0
- False Negatives: 55

C. Class Imbalance

The dataset used for Then project contains a significant class imbalance, with a vast majority of non-fraudulent job postings (approximately 95.6%) and a smaller number of fraudulent job postings (approximately 4.4%). Then class imbalance can pose challenges for predictive modeling, as it may lead to biased model performance.

VII. CONCLUSION

The development and implementation of the fraudulent job posting detection system have provided valuable insights and practical solutions to enhance the integrity and safety of online job markets. The system has effectively tackled the critical issue of fraudulent job advertisements, which pose substantial risks to job seekers in terms of scams, identity theft, and financial fraud. Through a systematic approach involving data collection, preprocessing, feature engineering, and machine learning modeling, a robust framework has been established to identify potentially fraudulent job postings successfully. The use of diverse machine learning algorithms, such as Decision Trees, Random Forests, and Support Vector Machines, has demonstrated promising results in distinguishing between authentic and deceptive listings. The evaluation of model performance, utilizing metrics like accuracy, F1-score, precision, and recall, has provided a comprehensive understanding of the system's capabilities. It has been recognized that addressing the class imbalance issue within the dataset through oversampling, undersampling, or synthetic data generation is crucial for achieving optimal results.The deployment of the system into a real-time production environment, equipped with a user interface and alerting system, ensures its practical utility for users and administrators. Swift assessments of job postings enable timely actions, further protecting job seekers from potential fraud. Furthermore, the incorporation of a feedback loop allows for continuous improvement, with user feedback and performance data guiding periodic model retraining and system enhancements.

The maintenance of a database for logging and reporting facilitates systematic monitoring and auditing, ensuring the system's ongoing effectiveness and trustworthiness over time. In conclusion, the fraudulent job posting detection system represents a comprehensive and proactive approach to address the challenges posed by deceptive listings in online job markets, contributing to a safer and more reliable environment for job seekers.

REFERENCES

[1] P. Kaur, “E-recruitment: a conceptual study,” International Journal of Applied Research, vol. 1, no. 8, pp. 78–82, 2015.

[2] C. Wang, “False information analysis of online recruitment,” Modern Marketing (Business Edition), vol. No. 335, no. 11, pp. 140-141, 2020

[3] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, “ORFDetector: ensemble learning based online recruitment fraud detection,” in 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1–5, Noida, India, 2019, August

[4] P. Rubin, “Regulation of information and advertising,” CPI Journal, vol. 4, 2008

[5] Y. Jiang, “On the construction of advertising credit supervision system,” Industrial and Commercial Administration, vol. 4, 2004

[6] S.Vidros, C. Kolias, and G. Kambourakis, “Online recruitment services: another playground for fraudsters,” Computer Fraud & Security, vol. 2016, no. 3, pp. 8–13, 2016.

[7] I.Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, “An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages,” in Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pp. 160–167, Athens, Greece, 2000, July.

[8] S.Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic detection of online recruitment frauds: characteristics, methods, and a public dataset,” Future Internet, vol. 9, no. 1, p. 6, 2017. [9] S. Dutta and S. K. Bandyopadhyay, “Fake job recruitment detection using machine learning approach,” International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, 2020

[9] O. Nindyati and I. G. B. B. Nugraha, “Detecting scam in online job vacancy using behavioral features extraction,” in 2019 International Conference on ICT for Smart Society (ICISS), vol. 7, pp. 1–4, Bandung, Indonesia, 2019, November

[10] S. Mahbub and E. Pardede, “Using contextual features for online recruitment fraud detection,” in the 27th International Conference on Information Systems Development, Lund, Sweden, 2018

[11] B. Alghamdi and F. Alharby, “An intelligent model for online recruitment fraud detection,” Journal of Information Security, vol. 10, no. 3, pp. 155–176, 2019.

[12] A. Mehboob and M. S. I. Malik, “Smart fraud detection framework for job recruitments,” Arabian Journal for Science and Engineering, vol. 46, no. 4, pp. 3067–3078, 2021.

[13] J. Kim, H. J. Kim, and H. Kim, “Fraud detection for job placement using hierarchical clusters-based deep neural networks,” Applied Intelligence, vol. 49, no. 8, pp. 2842–2861, 2019.

[14] G. Salton and C. T. Yu, “On the construction of effective vocabularies for information retrieval,” ACM SIGPLAN Notices, vol. 10, no. 1, pp. 48–60, 1975.