Real And Fake Job Post Detection Using Machine Learning

Mr. K. Hari Veerraju[1], A. Saranya[2], D. Prasanna[3], G. Mounika[4] , D. Karthikeya[5]

[1] Assistant Professor, Department of Computer Science
[2-5] B.Tech Student, Department of Computer Science
[1-5] Raghu Engineering College, Visakhapatnam

---***---

Abstract - The transition to online platforms has streamlined processes, reducing manual efforts. Job postings, now predominantly online, offer companies a broader reach for talent acquisition. However, amidst legitimate postings, fraudulent ones exist. This study aims to distinguish between real and fake job postings using machine learning techniques with optimal accuracy. Employing various data mining methods and classification algorithms—Logistic Regression, KNN, Decision Tree, XGBoost, Support Vector, Random Forest, and Multilayer Perceptron—we predict job authenticity. Supervised machine learning techniques guide dataset analysis, including variable identification and univariate, bivariate, and multivariate analyses, alongside handling missing values. Comprehensive data validation, cleaning, preparation, and visualization are conducted. Our experimentation, utilizing the Employment Scam Aegean Dataset (EMSCAD) with 17,881 samples, informs our approach. Ultimately, this research aims to enhance job seekers' security by effectively identifying fraudulent job postings. remove the writing issues.

Key Words: Logistic Regression, KNN, Decision Tree, XGBoost, Support Vector, Random Forest, and Multilayer Perceptron.

1. INTRODUCTION

Current data-driven approaches for detecting fake news typically rely on extracting credibility-indicative features from relevant articles, such as skeptical viewpoints and conflicting opinions. However, these methods face several limitations. Firstly, obtaining fake news is challenging, resulting in small datasets that hinder model effectiveness. Additionally, a significant portion of unverified news lacks contrasting perspectives, complicating credibility assessment. Furthermore, the disparity between true and false news extends beyond conflict features, including linguistic nuances like exaggerated emotional expression in fake news or sensationalist writing styles in clickbait. Consequently, existing approaches struggle to comprehensively capture these differences. Overcoming these limitations requires innovative strategies for data collection and feature extraction to discern the multifaceted nature of fake news, enhancing detection accuracy and reliability.

2. LITERATURE SURVEY

“Predicting of Job Failure in Compute Cloud Based on Online Extreme Learning Machine: A Comparative Study by Chunhong Liu1, 2, Jingjing Han2,Yanlei, Chuanchang Liu1, Bo Cheng1, and Junliang Chen1 in 2017”: [1]
The paper proposes a novel method, utilizing an Online Sequential Extreme Learning Machine (OS-ELM), for predicting job termination status in large-scale data centers. Unlike traditional offline methods, this approach enables real-time prediction as data arrive sequentially, enhancing resource utilization efficiency.

“Fake Job Recruitment Detection Using Machine Learning Approach by Samir Bandyopadhyay, Shawni Dutta in 2020”: [2]
The study proposes an automated program that uses machine learning-based categorization approaches to prevent fraudulent job ads on the internet. The best employment scam detection model is found by comparing the output of several classifiers that are used to verify fraudulent posts on the internet.

“Machine Learning and Job Posting Classification: A Comparative Study by Ibrahim M. Nasser1 and Amjad H. Alzaanin2 in 2020”: [3]
In this research, they looked at a text classification problem using several machine learning classifiers. Real and false job posts are included in the data that we used. The data was cleaned and preprocessed, and TF-IDF was then used to extract features.

“Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset by Sokratis Vidros 1, Constantinos Kolias 2, Georgios Kambourakis 1,2 and Leman Akoglu in 2017”: [4]
The transition to cloud-based hiring systems has introduced risks like online recruitment frauds (ORF), particularly employment scams, which have been largely unaddressed until now. This work defines ORF as a significant cybersecurity concern.

“Enhanced RSA Algorithm using Fake Modulus and Fake Public Key Exponent by Raghunandhan K R, Ganesh Aithal, Surendra Shetty, Rakshith N in 2018” :[5]
To strengthen the security of public key components which are essential for protecting data during communication the study presents an improved RSA algorithm. The suggested method replaces the normal public key components - modulus

n and exponent e - in classic RSA with fictitious public key exponent f and modulus X, respectively. By doing so, the algorithm aims to increase the complexity of factoring, thus mitigating the risk of integer factorization attacks and enhancing overall security.

3. PROPOSED SYSTEM

Below is a proposed machine learning-based real and fake job post detection:

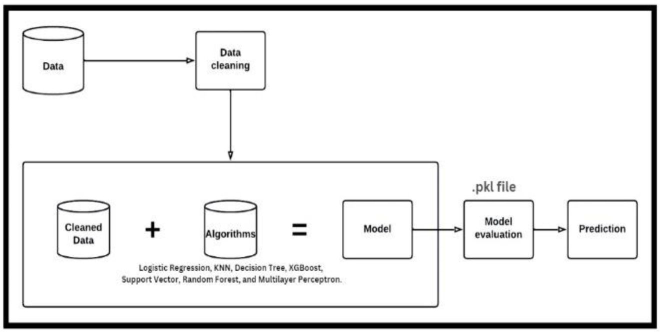

Data
Data cleaning
.pkl file
Cleaned Data
+
Algorithms
=
Model
Model evaluation
Prediction
Logistic Regression, KNN, Decision Tree, XGBoost, Support Vector, Random Forest, and Multilayer Perceptron.

Fig 1: Block Diagram of System

Proposed System: To solve this difficulty, the suggested strategy involves developing a machine learning model to identify genuine fraudulent job advertisements. Preprocessing is applied to the dataset, and columns are examined to determine whether variables are dependent or independent. Afterward, diverse machine learning techniques are implemented to identify trends and attain optimal precision in the outcomes.

Exploratory Data Analysis: To create a complete dataset, several datasets from various sources are combined. Afterward, multiple algorithms for machine learning are employed to identify patterns and get the highest level of precision in the outcomes.

Data Wrangling: This section involves loading the data, assessing its cleanliness, and then trimming and cleaning the dataset for analysis. Each cleaning decision is meticulously documented and justified.

Data Collection: To identify phony job postings, we employed EMSCAD. The dataset contains 17881 records and a total of 18 attributes per row.

Building the Classification Model: When used for classification difficulties, this method produces better results. It manages a mixture of discrete, categorical, and continuous data as well as outliers and irrelevant variables with effectiveness. It also yields an out-of-bag estimation error, which is impartial in several tests.

Algorithms: The cleansed data is utilized for using machine-learning algorithms after feature selection and cleaning. KNN, MultiLayer Perceptron, Support vector, XGBoost, Decision Tree, Random Forest, and Logistic Regression.

4. METHODOLOGY

Objectives:

By comparing supervised algorithms, the goal is to develop a machine-learning model that can anticipate legitimate vs fraudulent jobs, perhaps displacing the updatable machine-learning classifier models.

Project Objectives:

A. Study of the exploratory data for variable identification:

1) Opening the prepared dataset.
2) Import all the required libraries.
3) Examine the normal characteristics.
4) Locate missing and duplicate values.
5) Verify unique and count values.

B. Analysis of univariate data

1) To change the name, add, or remove data.
2) To indicate the datatype.

C. Analyzing exploration data in two or more variables

1) heatmap, pairplot, barchart, and also using histplot diagrams.

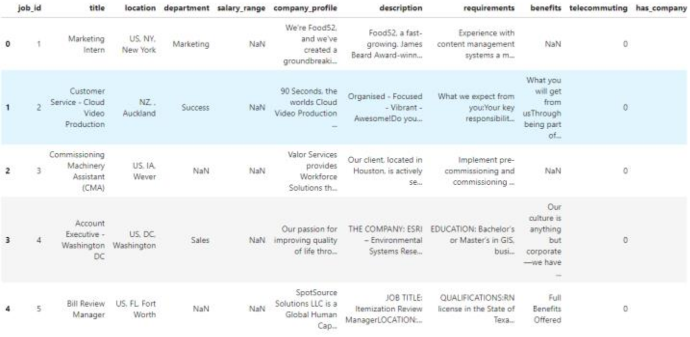
D. Utilizing characteristic engineering to identify anomalies

1) getting the dataset ready for analysis.
2) separating the training and test sets
3) contrasting random forests, logistic regression models, decision trees, etc.

5. RESULT & ANALYSIS

A. Loading Dataset:

This dataset consists of 17,881 records (samples) of data. The attributes are “job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, telecommunication, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent”.



	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	0	
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	0	
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	

Fig 2: Dataset

B. Cleaning and Distribution of Dataset

WordCloud of job titles: Fig 4 shows the visual representation of unstructured data, which in this case are the job titles.

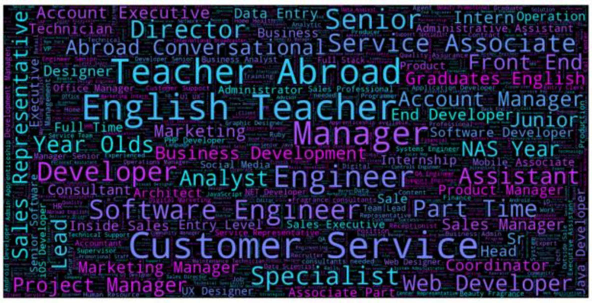


Fig 3: WordCloud of Job Titles

The distribution of the target feature (fraudulent): Fig 4 shows the normal barplot distribution of the fraudulent variable which is of 17880 rows in the dataset.

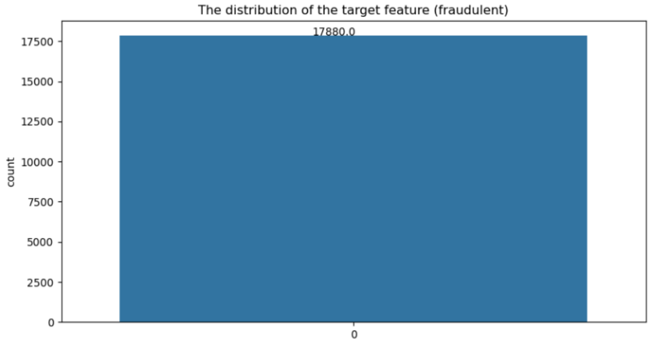


Fig 4: Fraudulent Distribution

Oversampling Target Variable: A method for addressing the class imbalance in statistical modeling and machine learning is called oversampling the target variable. When one class (the minority class) in a classification problem is substantially underrepresented in comparison to the other class(es) (the majority class or classes), this is referred to as class imbalance.

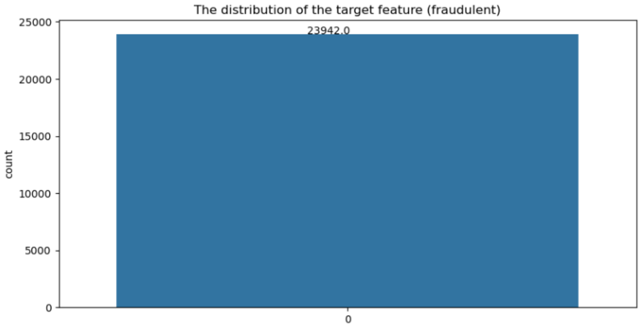


Fig 5: Oversampling of target variable

For this reason, the fraudulent variable is oversampled by adding redundancy to it, which made its size from 17880 to 23942 as shown in Fig 5.

C. Accuracy of the Algorithms

Logistic Regression: By using logistic regression, as shown in Fig 6, we obtained an accuracy of 80.79%.

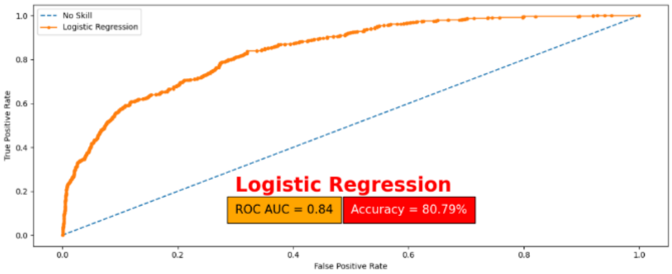


Fig 6: Logistic Regression

Support Vector Classifier: By using Support Vector Classifier, as shown in Fig 7, we have obtained an accuracy of 87.76% in finding real and fake job posts.

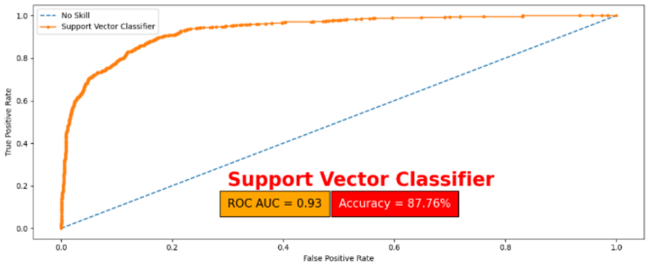


Fig 7: Support Vector Classifier

MultiLayer Perceptron Classifier: From Fig 8, we can observe that, by using MultiLayer Perceptron Classifier which is one of the machine learning classification algorithms we have obtained 91.98% of classifying real and fake job posts.

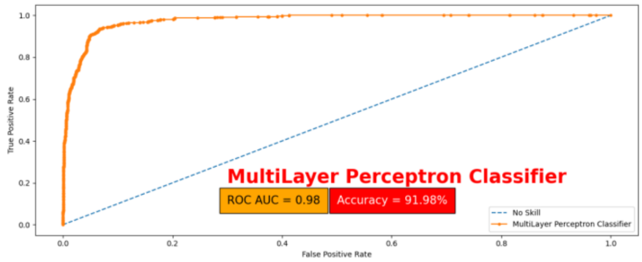


Fig 8: MultiLayer Perceptron Classifier

KNN Classifier: The KNN classifier has been used which is most widely used for its cost-effectiveness, and easy-to-use. From Fig 9, we observed that it has achieved 96.1% accuracy in classifying real and fake job postings.

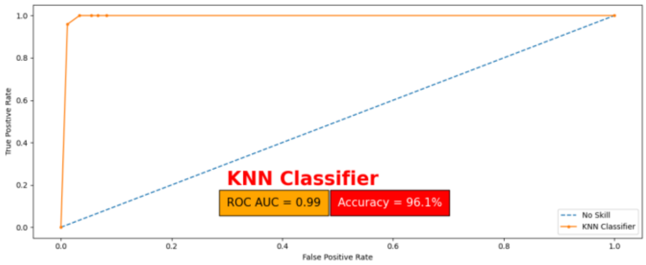


Fig 9: KNN Classifier

Decision Tree Classifier: From Fig 10, we can observe that Decision Tree Classifier has achieved an accuracy of 98.43% when applied on the dataset.

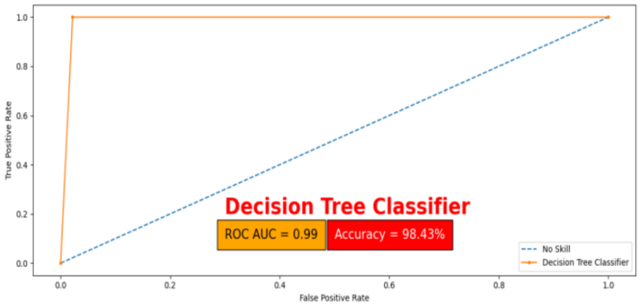


Fig 10: Decision Tree Classifier

XGBoost Classifier: The XGBoost Classifier has achieved an accuracy of 99.35% for the classification of real jobs from fake jobs in the dataset, as shown in Fig 11.

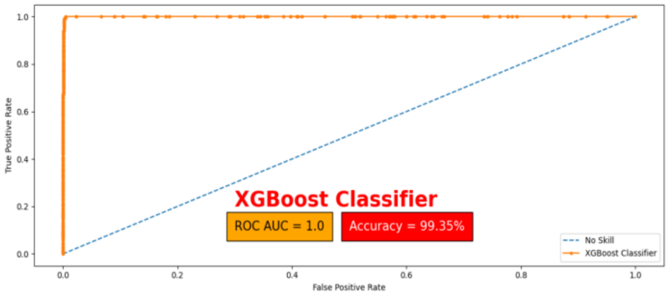


Fig 11: XGBoost Classifier

Random Forest classifier: By using the Random Forest Classifier on the dataset, we can observe that, from Fig 12, it scored 99.83% accuracy.

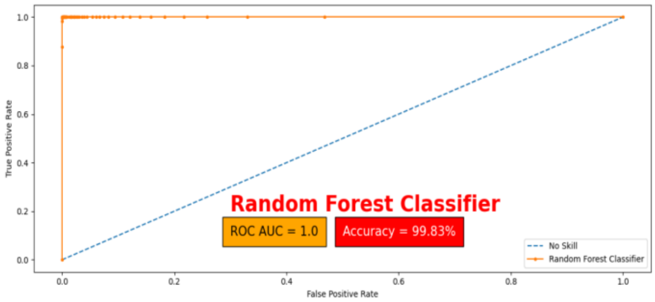


Fig 12: Random Forest Classifier

D. Comparision among the classifiers: Fig 13, shows the plotting curve of all the classifiers on the x-axis and their accuracy along the y-axis. Starting from Logistic regression to Random Forest Classifier.

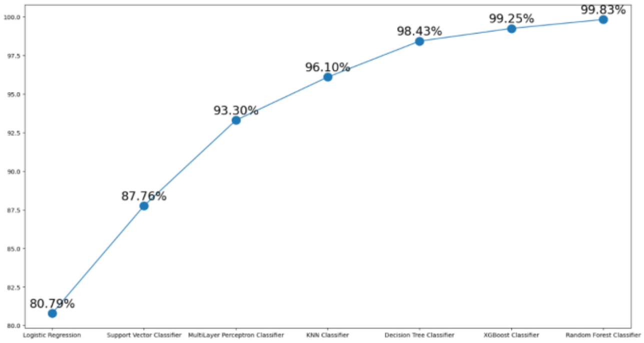
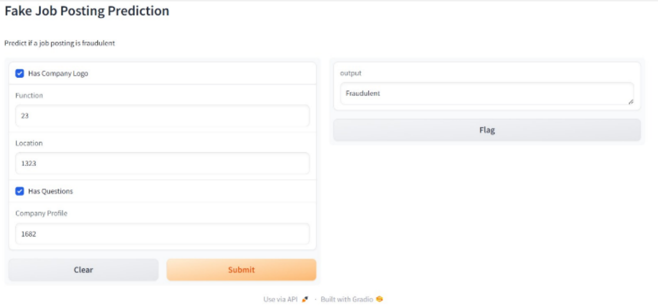


Fig 13: Classifiers Accuracy Plot

Project Result: The below figure shows the outcome of the project, the interface is implemented using the Gradio module of Python. The input parameters are Has company logo, function, location, questions, and company profile.



Fake Job Posting Prediction

Predict if a job posting is fraudulent

☑ Has Company Logo

Function: 23

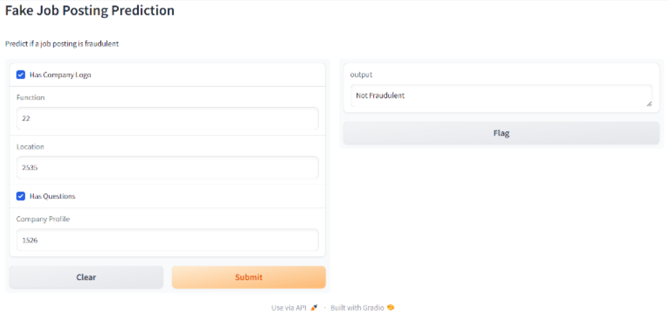
Location: 1323

☑ Has Questions

Company Profile: 1682

output: Fraudulent

Fig 14: Fake Job Post Prediction (Fraudulent)



Fake Job Posting Prediction

Predict if a job posting is fraudulent

☑ Has Company Logo

Function: 22

Location: 2535

☑ Has Questions

Company Profile: 1526

output: Not Fraudulent

Fig 15: Fake Job Post Prediction (Not Fraudulent)

For these inputs, the trained model of the XGBoost classification algorithm is applied and classifies whether the job is real or fake.

6. CONCLUSION

After data has been cleaned and processed, missing values are addressed, exploratory analysis is carried out, and model construction and assessment are the final steps in the analytical process. Getting the best accuracy score on the public test set is the aim. This tool helps distinguish between legitimate and fraudulent job advertisements. It approaches each step methodically to make sure the data is clean and ready for analysis.

To preserve data integrity, data must first go through a thorough cleaning process in which any mistakes or inconsistencies are fixed. Missing values are then carefully managed to avoid any distortion or bias in the analysis. Investigative analysis explores the subtleties of the dataset, revealing trends and revelations that guide further modeling choices. Choosing and training suitable algorithms for the particular purpose of differentiating between real and fake job advertising is known as model development. Metrics for evaluation assess the model's performance and allow for iterative refinement to improve accuracy and dependability.

In this project, we have achieved 99.83% accuracy with the Random Forest Classifier, 99.25% accuracy with the XGBoost Classifier, 98.43% accuracy with the Decision Tree Classifier, 96.10% accuracy with the KNN classifier, 93.30% accuracy with the MultiLayer Perceptron Classifier, 87.76% accuracy with Support Vector Classifier, 80.79% accuracy with Logistic Regression Classifier.

REFERENCES

1. “Predicting of Job Failure in Compute Cloud Based on Online Extreme Learning Machine: A Comparative Study by CHUNHONG LIU1, 2, JINGJING HAN2, YANLEI SHANG1, CHUANCHANG LIU1, BO CHENG1, AND JUNLIANG CHEN1 in 2017” [1]
2. “Fake Job Recruitment Detection Using Machine Learning Approach by Samir Bandyopadhyay, Shawni Dutta in 2020” [2]
3. “Machine Learning and Job Posting Classification: A Comparative Study by Ibrahim M. Nasser1 and Amjad H. Alzaanin2 in 2020” [3]
4. “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset by Sokratis Vidros 1, Constantinos Kolias 2, Georgios Kambourakis 1,2 and Leman Akoglu in 2017” [4]
5. “Enhanced RSA Algorithm using Fake Modulus and Fake Public Key Exponent by Raghunandhan K R, Ganesh Aithal, Surendra Shetty, Rakshith N in 2018” [5]
6. S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006. [6]