Unmasking Deception: A Proactive Approach to Detecting and Preventing Fake Job Offers

Lembhe Akshata[1], Sapkal Dipraj[2], Khairnar Gaurav [3]

[1]Assistant Professor, Department of Statistics, Dr. D. Y. Patil ACS College Pimpri Pune 18
[2,3]Student, Department of Statistics, Dr. D. Y. Patil ACS College Pimpri Pune 18

Abstract

The research presents an innovative approach to combating fraudulent job postings on the internet through the utilization of machine learning-based classification techniques. By leveraging different classifiers, including single classifiers and ensemble classifiers, the study aims to discern fake job postings amidst a vast array of online listings. Through an extensive analysis of experimental results, the research identifies ensemble classifiers as the optimal choice for detecting employment scams, surpassing the efficacy of single classifiers.

Employing a dataset sourced from Kaggle, the study focuses on distinguishing between real and fake job postings, with the latter comprising a minority of the dataset, as anticipated.

By adhering to these structured stages, the research aims to contribute to the advancement of methods for identifying and mitigating fraudulent activities in online job postings, thereby enhancing the integrity of online recruitment processes.

Keywords: Attrition, Classifier, algorithms

Problem Statement

The objective of this project is to develop a classifier capable of distinguishing between genuine and fraudulent job postings. The evaluation of the classifier's performance will be based on two distinct models tailored for numeric and textual data features, respectively. The final output will integrate the outcomes of both models to provide a comprehensive assessment of job authenticity.

Metrics

Two primary metrics will be utilized to evaluate the models:

Accuracy: This metric is defined by this formula

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{Truenegative}}{\text{TruePositive} + \text{Truenegative} + \text{FalsePositive} + \text{FalseNegetive¿¿}}$$

This metric gauges the proportion of correctly classified data points relative to the total number of data points. While accuracy provides an overall assessment of classification performance, it may be influenced by imbalanced class distributions. In the context of identifying both real and fake jobs, an accurate classification of both categories is essential.

F1-Score: F1 score is a measure of a model’s accuracy on a dataset.

The formula for this metric is –

F1 =

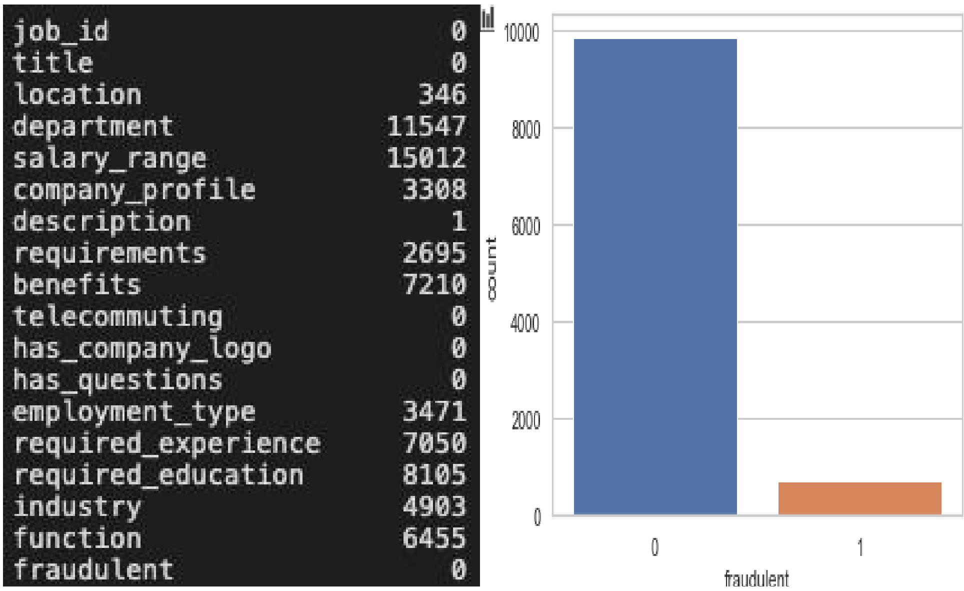
The F1 score, a composite metric of precision and recall, offers a balanced evaluation of a model's accuracy. It considers false positives and false negatives, which are particularly relevant in scenarios where misclassifying either category can have significant consequences. By incorporating both precision and recall, the F1 score provides a robust measure of a model's effectiveness in classifying both real and fake job postings.

These metrics collectively serve to assess the classifier's ability to accurately differentiate between real and fake job postings, accounting for the inherent challenges posed by imbalanced class distributions and the importance of minimizing both false positives and false negatives.

Analysis

Data Exploration

The data for this project is available at Kaggle - https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction.The dataset consists of 17,880 observations and 18 features.



```
job_id                     0
title                      0
location                 346
department             11547
salary_range           15012
company_profile         3308
description                1
requirements            2695
benefits                7210
telecommuting              0
has_company_logo           0
has_questions              0
employment_type         3471
required_experience     7050
required_education      8105
industry                4903
function                6455
fraudulent                 0
```

Figure 1

Given the predominant data types of Booleans and text, traditional summary statistics aren't pertinent to this analysis. The sole integer variable, job_id, is deemed irrelevant and thus omitted from further consideration. During dataset exploration, it becomes evident that variables like department and salary range contain numerous missing values and are consequently dropped from subsequent analysis.

Upon initial assessment, it's noted that the job postings span multiple languages due to their origin from various countries. To streamline the process, the project focuses solely on US-based postings, which represent approximately 60% of the dataset, ensuring consistency in language (English) for easier interpretation. Additionally, location data is refined, splitting it into state and city components for enhanced analysis. The resulting dataset comprises 10,593 observations and 20 features.

Notably, the dataset exhibits significant class imbalance, with 9,868 (93%) postings classified as genuine and only 725 (7%) identified as fraudulent. A visual representation, such as a count plot, vividly illustrates this disparity.

EXPLORATORY VISUALIZATION

The initial visualization of the dataset involves constructing a correlation matrix to examine the connections among numeric variables.

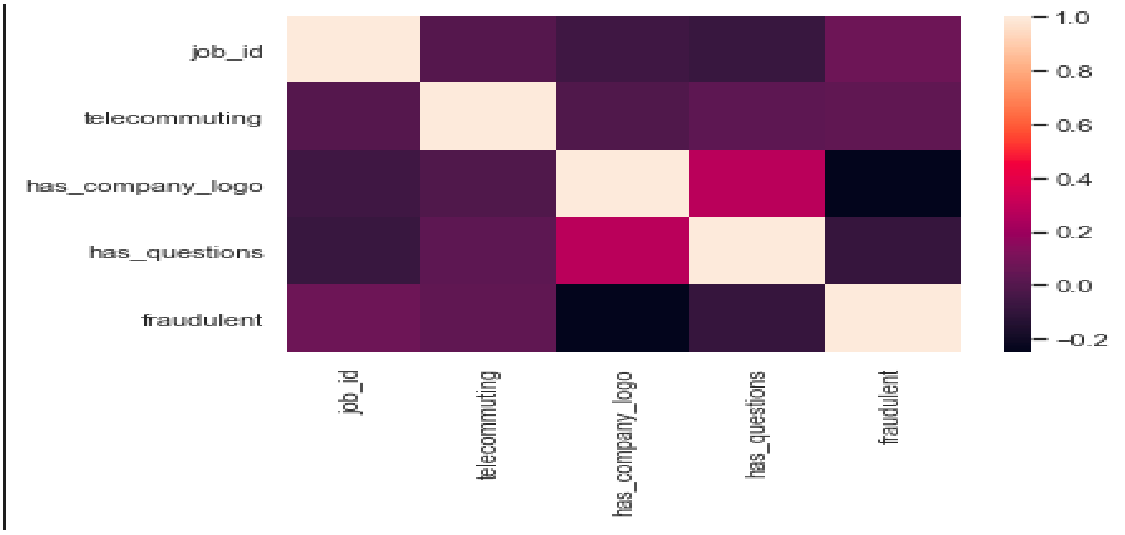


Figure 2

Surprisingly, no pronounced positive or negative correlations emerge within the numeric data. However, an intriguing pattern emerges regarding the Boolean variable "telecommuting." Notably, instances where this variable registers a value of zero are strongly associated with a 92% likelihood of the job being fraudulent.

Following the examination of numeric features, the focus shifts to exploring textual characteristics within the dataset, with the exploration commencing from the perspective of location data.

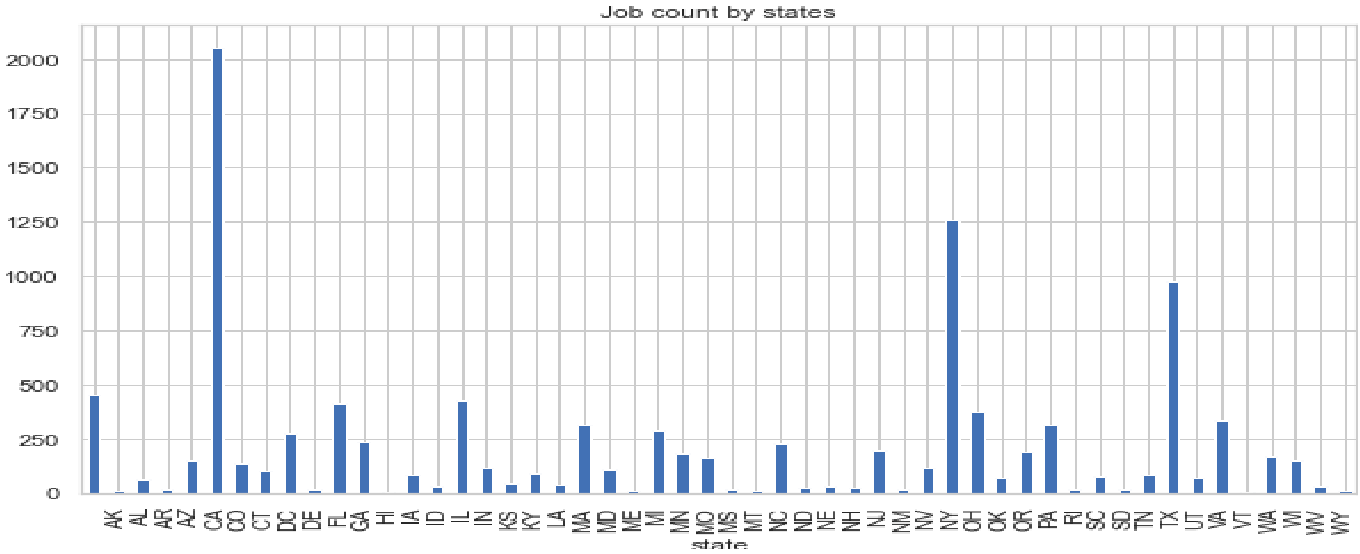


Figure 3

The visual representation depicts the states that generate the highest volume of job postings, with California, New York, and Texas emerging as the top contributors. This observation prompts further exploration through the creation of another bar plot, specifically focusing on the distribution of both fake and authentic job listings within the top 10 states.

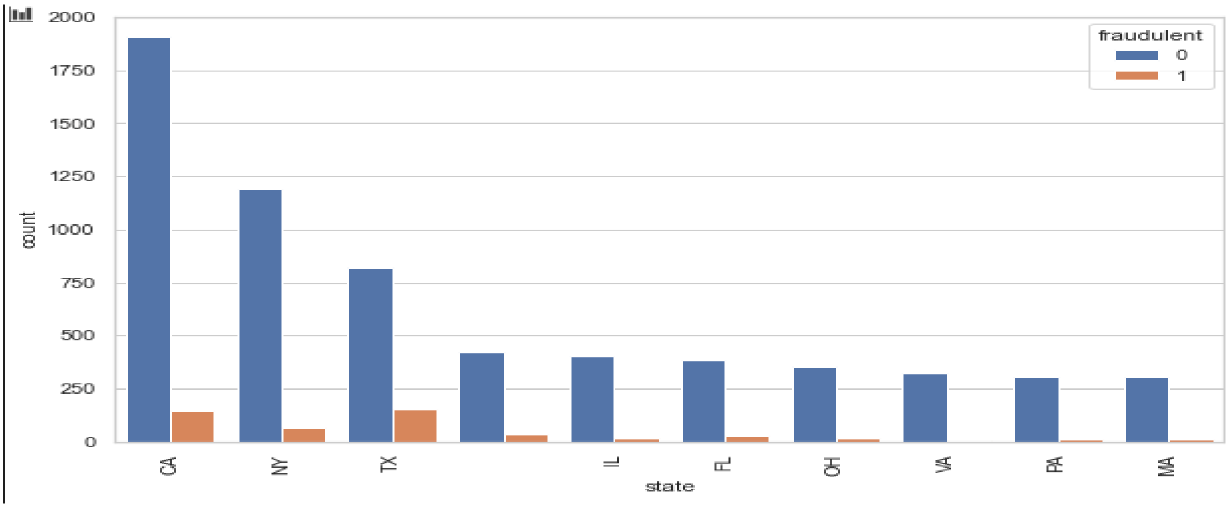


Figure 4

The depicted graph indicates a higher likelihood of encountering fraudulent job postings in Texas and California relative to other states.

To delve further into the analysis and incorporate both states and cities, a ratio is computed to gauge the prevalence of fake jobs relative to real ones. This ratio is calculated by dividing the count of fake jobs by the count of real jobs within each state and city combination, utilizing the formula:

$$\text{Ratio} = \frac{\text{State¿City¿Fraudulent¿1}}{\text{State¿City¿Fraudulent¿0}}$$

Only ratio values equal to or greater than one are considered for plotting purposes.

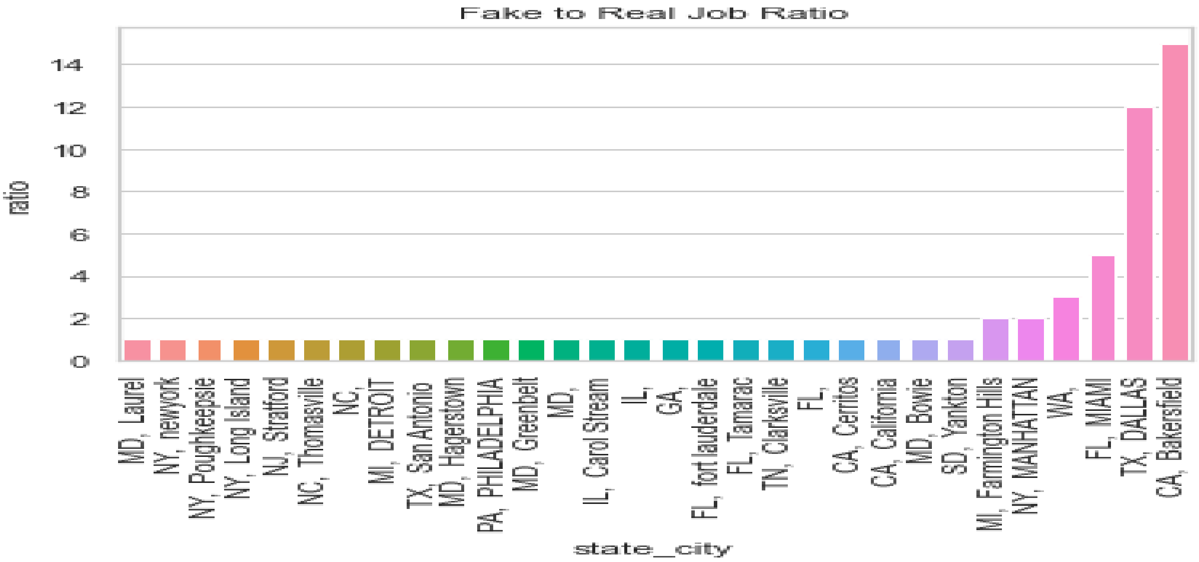


Figure 5

Bakersfield in California exhibits a notably high fake to real job ratio of 15:1, while Dallas, Texas follows closely with a ratio of 12:1. Given these ratios, job postings originating from these locations are likely to have a significantly elevated risk of being fraudulent. Further exploration delves into other text-based variables to uncover potential relationships of importance.

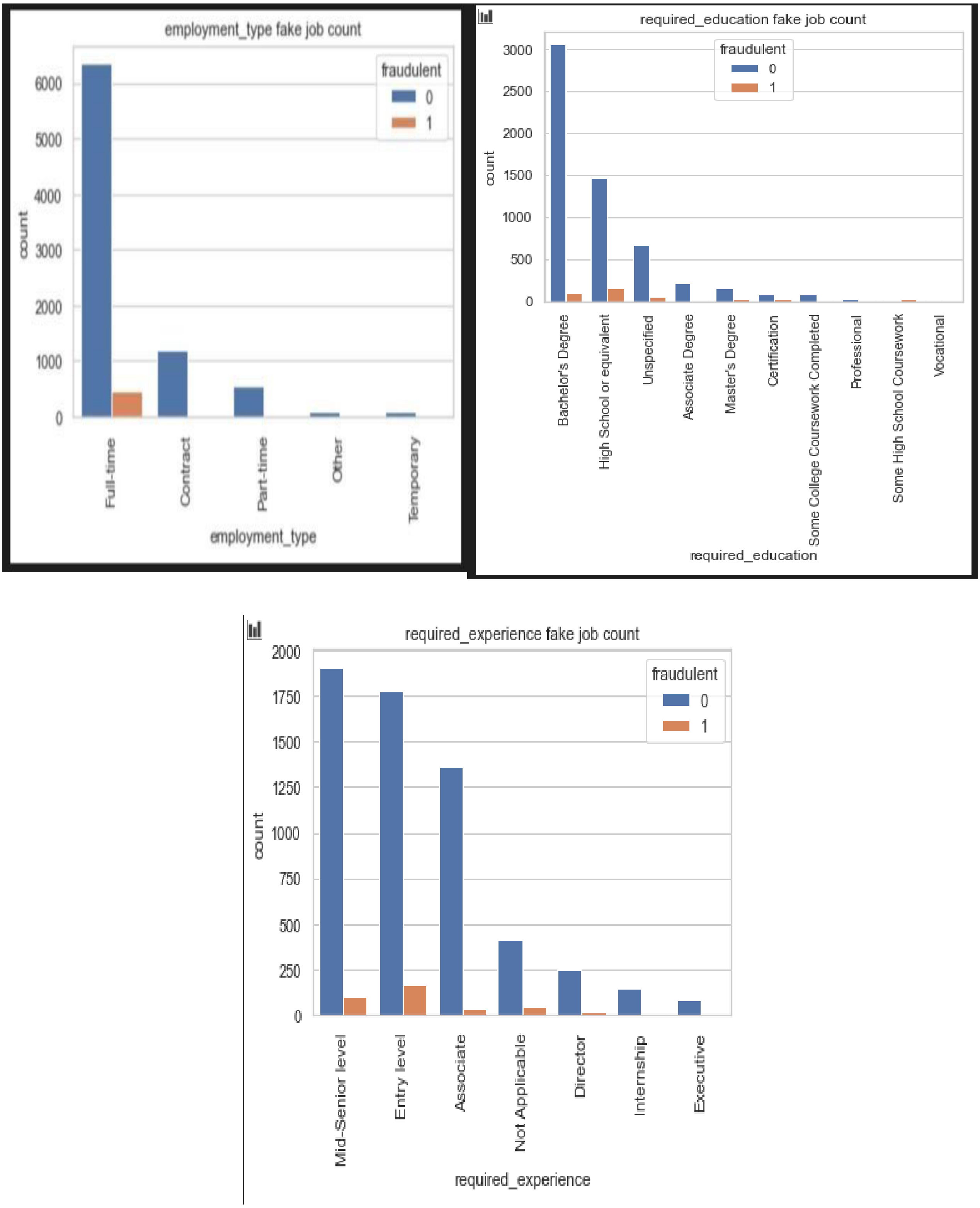


Figure 6

The visuals above illustrate a clear trend wherein the majority of fraudulent job listings fall under the full-time category. Additionally, these positions commonly target entry-level roles that necessitate either a bachelor's degree or a high school education.

Expanding the analysis on text-related fields, a consolidation is performed to merge various text-based categories into a single field labelled "text." This amalgamation encompasses fields such as title, location, company profile, description, requirements, benefits, required experience, required education, industry, and function. Subsequently, a histogram depicting character counts is examined to discern distinctions between real and fake job postings. Notably, while the character count distribution appears relatively similar for both real and fake jobs, real job postings exhibit a higher frequency overall.

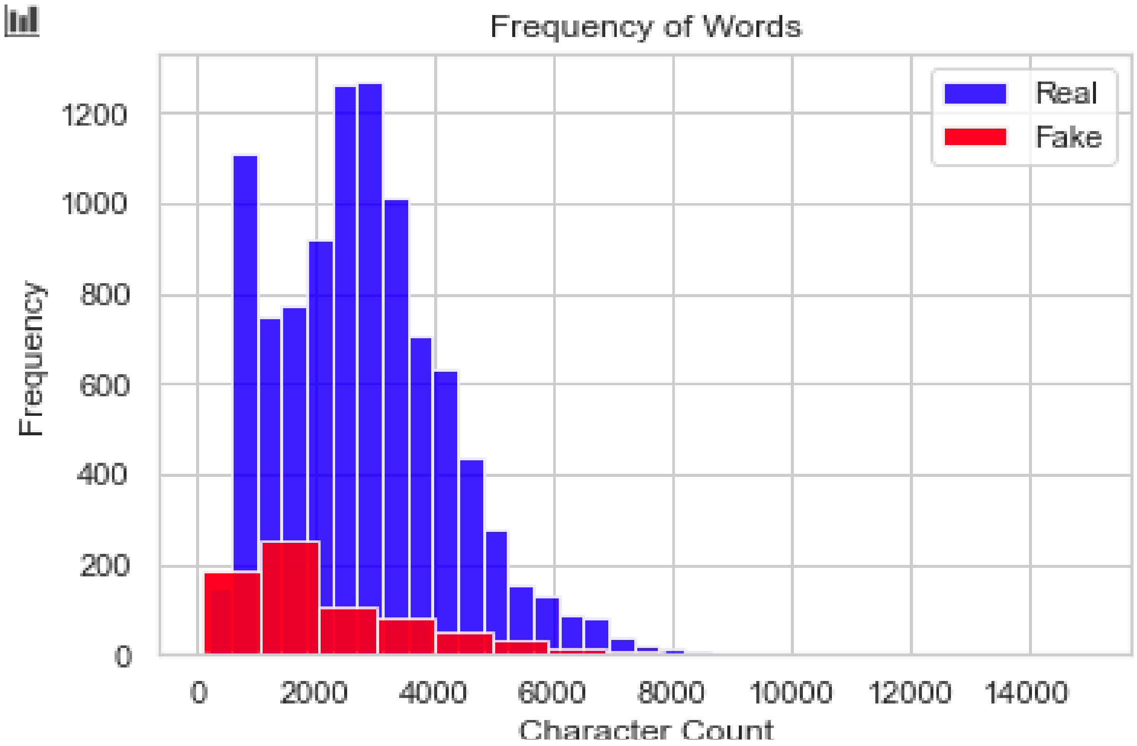


Figure 7

Graph depicting job counts categorized by employment type, required experience, and fraudulent status reveals notable trends. Specifically, it is apparent that full-time positions are more prevalent compared to other types of employment. Moreover, there appears to be a higher frequency of job listings demanding mid-senior level experience, indicating a significant presence of such opportunities within the dataset.

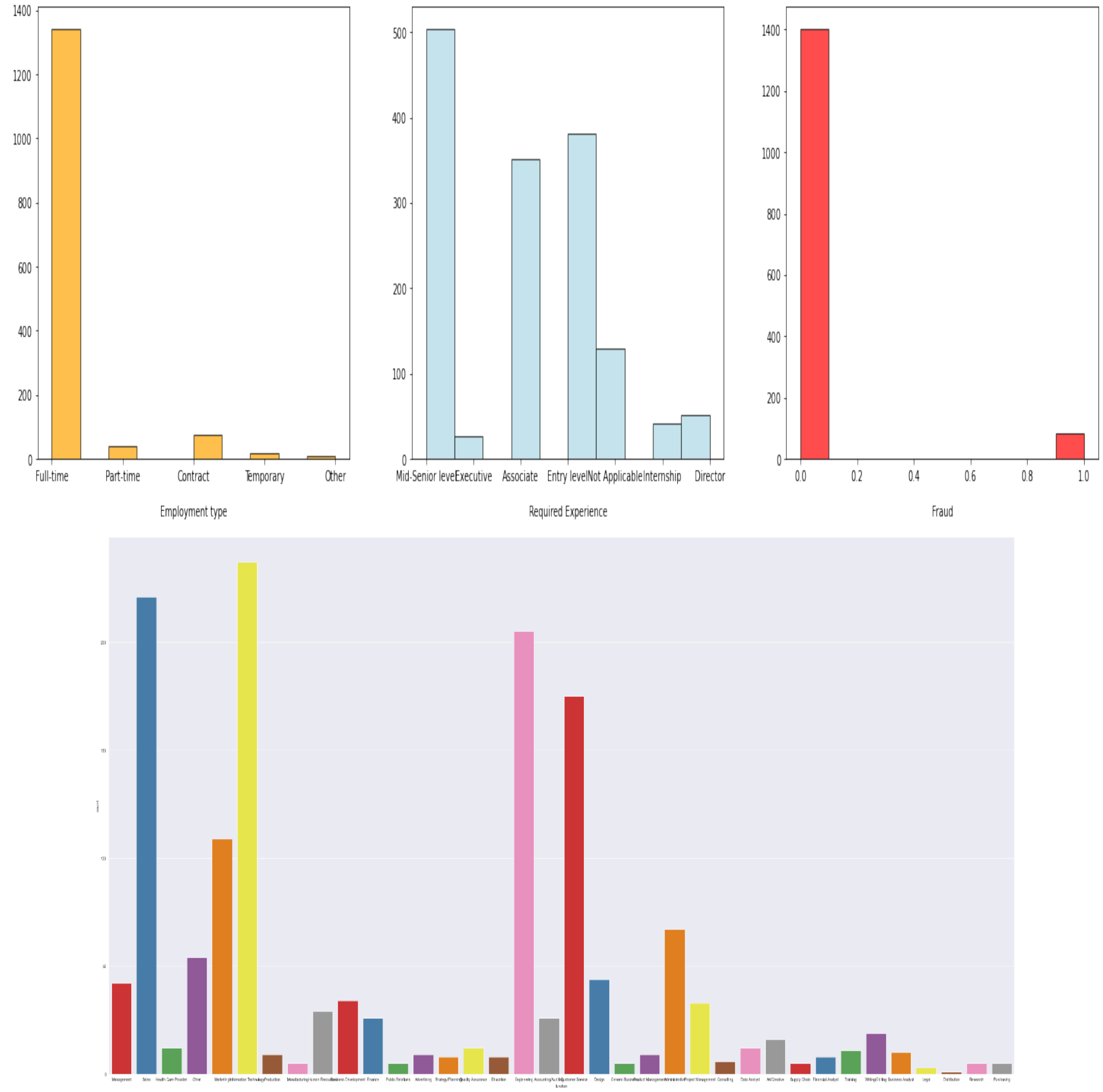


Figure 8

Analysing the graph illustrating job counts by function, it's evident that the IT sector boasts the highest volume of job postings, while Distribution registers the lowest count. Notably, roles within Sales, IT, Marketing, Engineering, Customer Service, and Administrative domains emerge as the most sought-after positions, indicating a strong demand for professionals in these fields.

ALGORITHMS AND TECHNIQUES

Based on the preliminary analysis, it's clear that both text and numeric data will be integral for the final modelling phase. A final dataset is curated with the following features for comprehensive analysis:

1. Telecommuting
2. Fraudulent status
3. Ratio: Fake to real job ratio based on location
4. Text: Consolidated information from title, location, company profile, description, requirements, benefits, required experience, required education, industry, and function

5. Character_count: Word count histogram of the textual data

These selected features will form the foundation for the subsequent modelling process, facilitating a thorough examination of the dataset.

Further pre-processing steps are essential before leveraging textual data for modelling. The project employs specific algorithms and techniques, including:

A. **Naive Bayes Classifier:** This classifier is a supervised learning tool rooted in Bayes' Theorem of Conditional Probability. Despite potentially inaccurate probability estimates, the decisions made by this classifier prove effective in practice. It excels when dealing with independent features or those that are functionally dependent. The classifier's accuracy is primarily influenced by the amount of information loss incurred due to the assumption of feature independence, rather than by the dependencies themselves.

B. **SGD Classifier:** The SGD Classifier operates as a linear classifier, utilizing optimization via Stochastic Gradient Descent (SGD). It distinguishes itself by leveraging SGD as an optimization technique, independent of the specific machine learning algorithm or model it's paired with, such as logistic regression or linear Support Vector Machine (SVM)

C. Multiple logistic regression is employed when analyzing a scenario involving one nominal variable and two or more measurement variables. Its purpose lies in determining how these measurement variables influence the nominal variable. This technique facilitates the prediction of probabilities associated with the dependent nominal variable. Alternatively, with careful consideration, it can offer insights into which independent variables exert significant influence on the dependent variable, aiding in decision-making processes.

D. **K-nearest Neighbour Classifier:** K-nearest Neighbour Classifiers, commonly referred to as lazy learners, classify objects by assessing their proximity to training examples within the feature space. This classifier operates by considering the k nearest objects when determining the class of a given sample. However, a key challenge associated with this classification technique lies in selecting the optimal value for k, which significantly impacts the model's performance.

E. **Decision Tree Classifier:** The Decision Tree (DT) classifier is characterized by its tree-like structure, representing a method for organizing knowledge on classification tasks. In this structure, each target class is represented as a leaf node, while non-leaf nodes serve as decision nodes, indicating specific tests to be performed. The outcomes of these tests determine the paths along the branches of the decision tree. Starting from the root node and traversing through the tree until a leaf node is reached, the classification result is obtained. Decision tree learning has found applications in various domains, including spam filtering, where it proves effective in forecasting outcomes based on defined criteria through model implementation and training.

F. **Random Forest:** Random Forest is a popular supervised machine learning algorithm utilized for both classification and regression tasks. It operates by constructing multiple decision trees on varied samples from the dataset and aggregates their predictions through a majority vote mechanism for classification problems or an averaging approach for regression tasks.

RESULTS

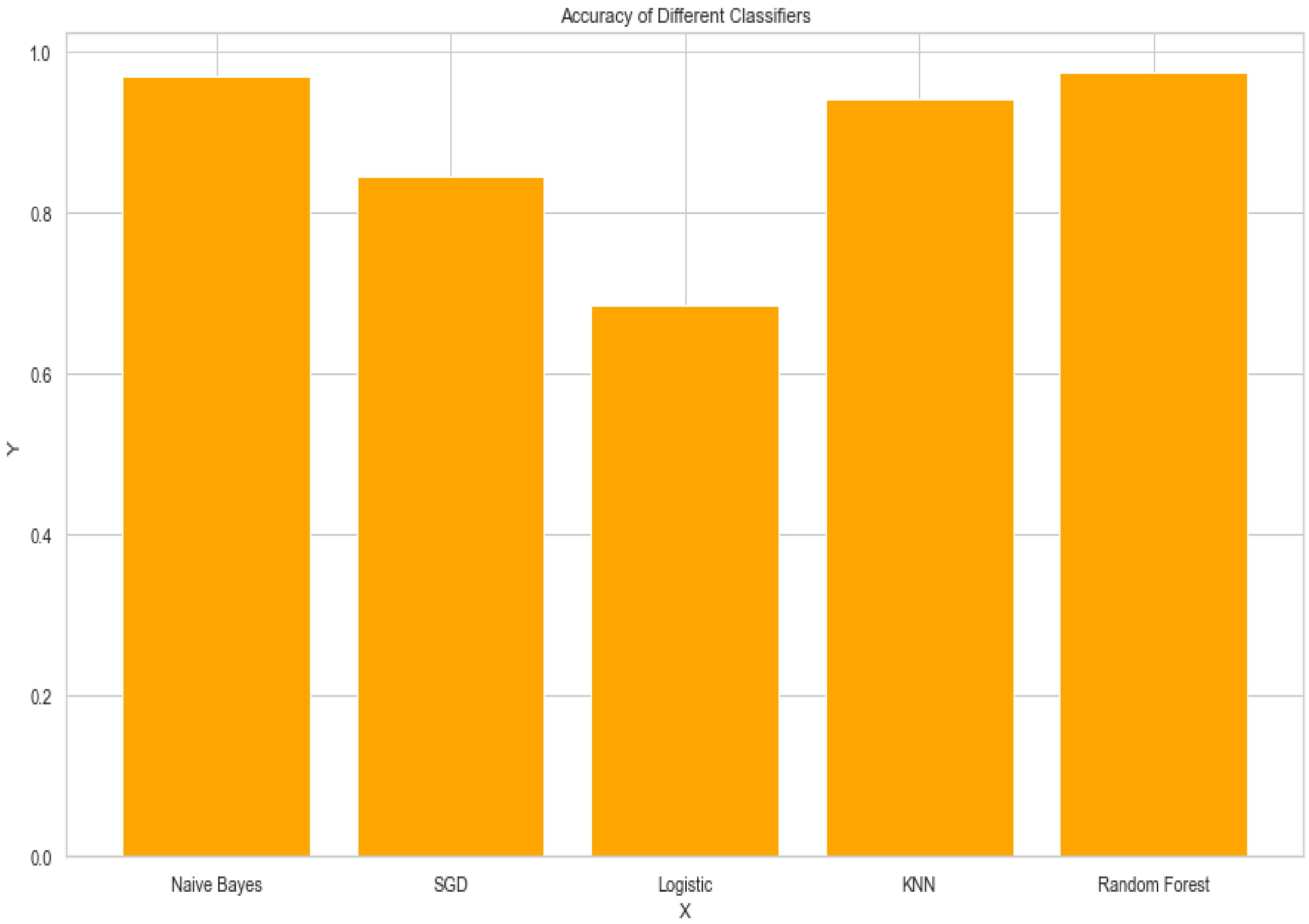
Model Evaluation and Validation: All the classifiers mentioned earlier are trained and tested on a dataset containing both fraudulent and legitimate job posts to detect fake postings. The comparative analysis presented in Table 1 showcases the performance metrics of each classifier. These metrics, including accuracy, f1-score, Cohen-kappa score, and Mean Squared Error (MSE), provide insights into the overall effectiveness of the classifiers in distinguishing between fake and genuine job postings.

TABLE I PERFORMANCE COMPARISON CHART FOR SINGLE CLASSIFIER BASED PREDICTION AND ENSEMBLE CLASSIFIER BASED PREDICTION

Performance Measure Metric	**Naïve Bayes Classifier**	**SGD Classifier**	**Logistic Regression Classifier**	**K-Nearest Neighbor Classifier**	**Random Forest Classifier**
Accuracy	97.13%	84.58%	68.66%	94.27%	97.48%
F1-Score	74.35%	82.13%	72.98%	63.09%	77.28%
Cohen Kappa Score	0.72	0.77	0.44	0.60	0.76
MSE	0.028	0.026	0.276	0.057	0.024

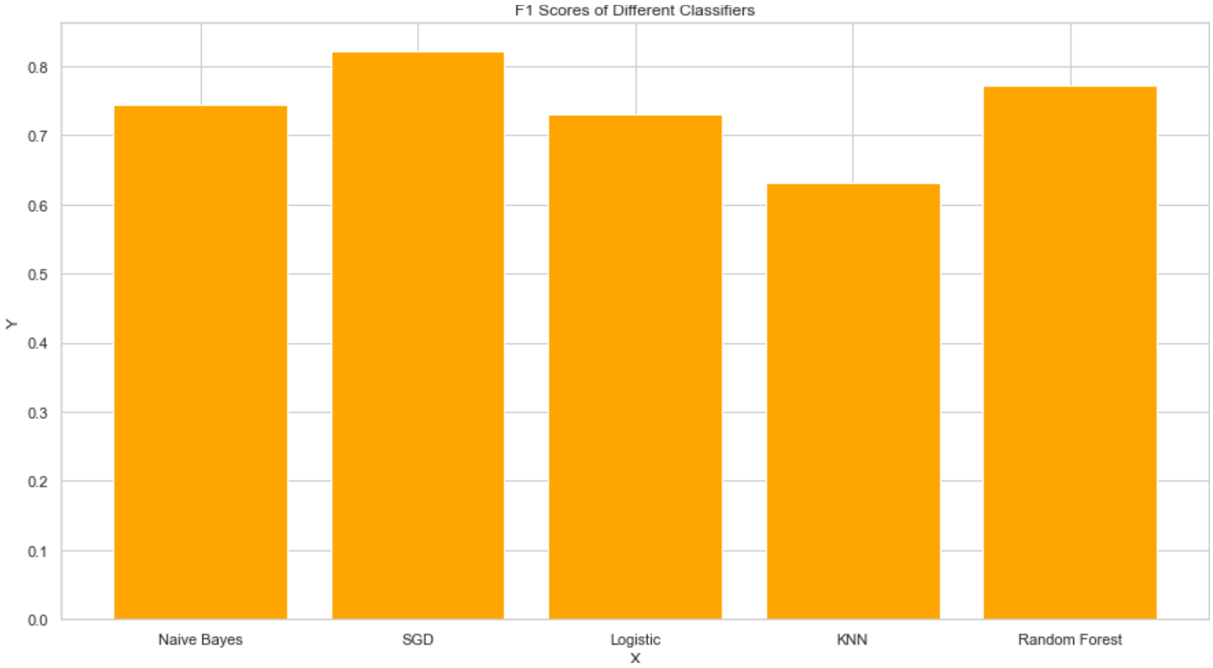
Based on the findings from Table 1, it's evident that the Random Forest Classifier outperforms the Naïve Bayes, Logistic Regression, K-Nearest Neighbor, and SGD Classifiers, displaying promising results in detecting fake job postings. As an ensemble classifier, the Random Forest Classifier proves to be a fruitful predictor, offering superior performance compared to its counterparts. Experimental results affirm that ensemble-based classifiers generally yield improved outcomes over other models specified in Table 1. While the Random Forest Classifier demonstrates a comparable F1-score to its competitors, it notably excels in other metrics, solidifying its position as the optimal model for this fake job detection scheme. sAfter implementing and comparing classifiers based on various metrics, experimental results have demonstrated that ensemble-based classifiers consistently deliver superior performance compared to other models listed in Table 1. Despite the Random Forest classifier yielding an F1-score similar to its competitors, it exhibits significant prowess in other evaluation metrics. Therefore, the Random Forest classifier emerges as the most effective model for detecting fake job postings within this scheme.

Accuracy of Different Classifiers:-



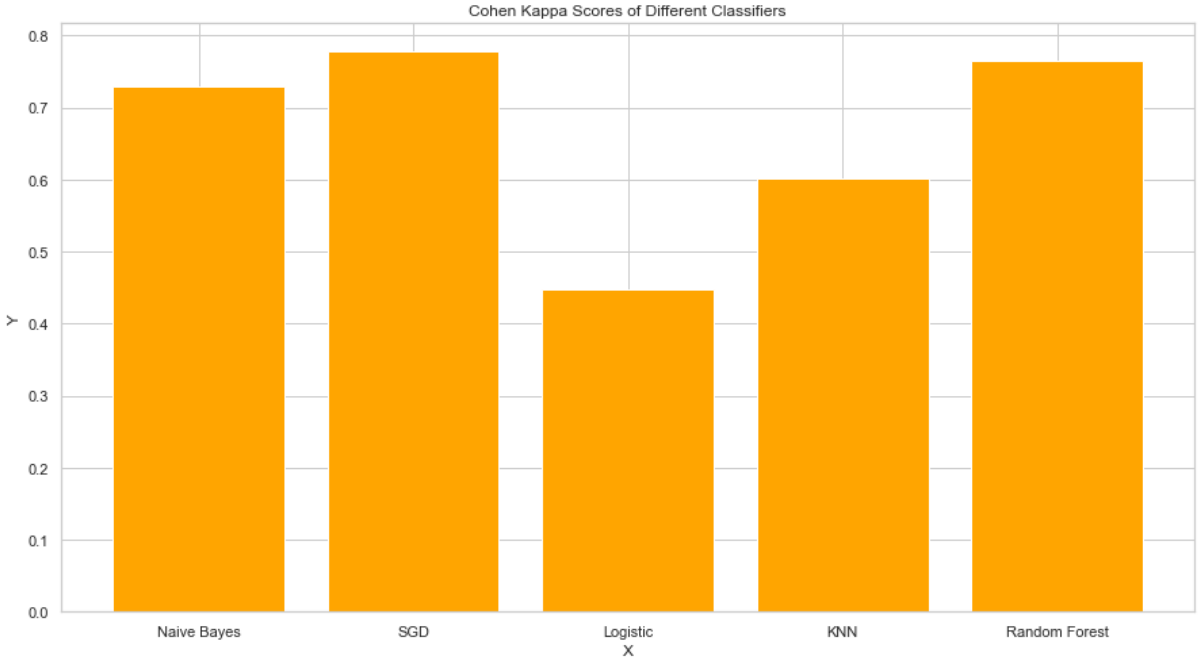
The analysis of the figure reveals that both the Naïve Bayes Classifier and Random Forest Classifier exhibit high accuracy scores, with their values being almost equal. Conversely, the accuracy of the Logistic Regression Classifier is notably lower. Consequently, for subsequent analysis, it is advisable to consider either the Naïve Bayes Classifier or the Random Forest Classifier due to their superior accuracy performance.

F1 score For the Different classifier



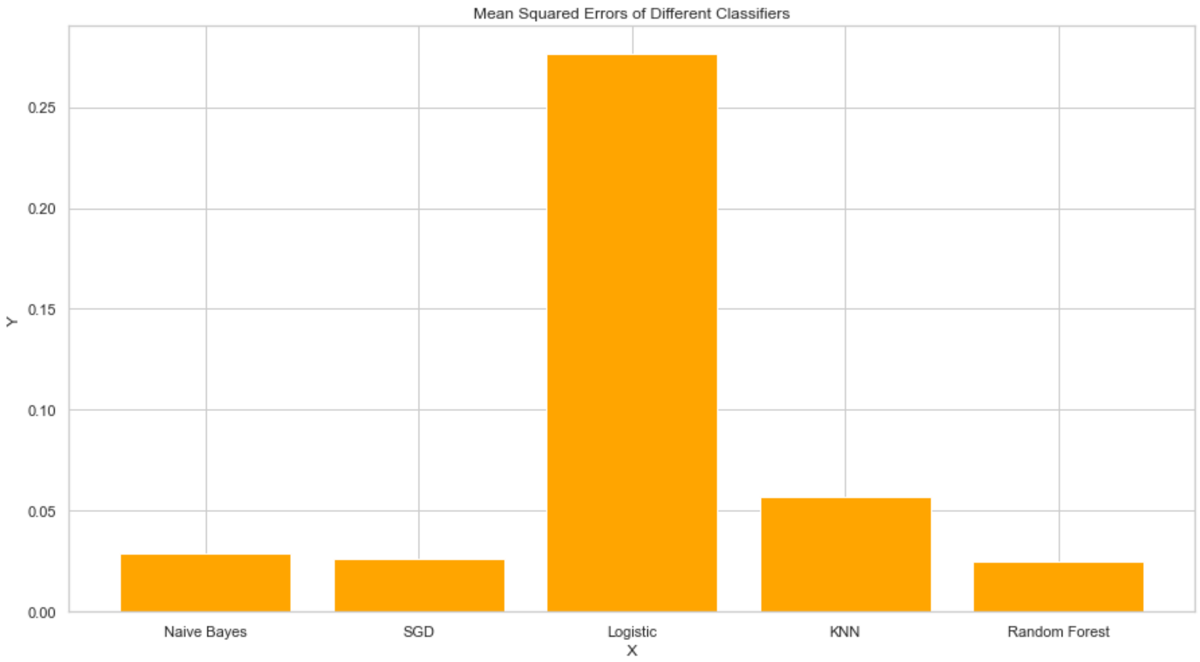
Upon examining the figure, it becomes apparent that both the SGD Classifier and Random Forest Classifier demonstrate high F1 scores. Conversely, the F1 score of the KNN Classifier is notably low. As a result, for subsequent analysis, it is advisable to consider either the SGD Classifier or the Random Forest Classifier due to their superior performance in terms of F1 score.

Cohen Kappa Score of Different Classifiers:-



Observing the figure, it's evident that both the SGD Classifier and Random Forest Classifier display high Cohen Kappa scores. Conversely, the Cohen Kappa score of the Logistic Regression model is notably low. Thus, for subsequent analysis, it is recommended to utilize either the SGD Classifier or the Random Forest Classifier due to their superior performance in terms of Cohen Kappa score.

Mean Square Error of Different Classifiers

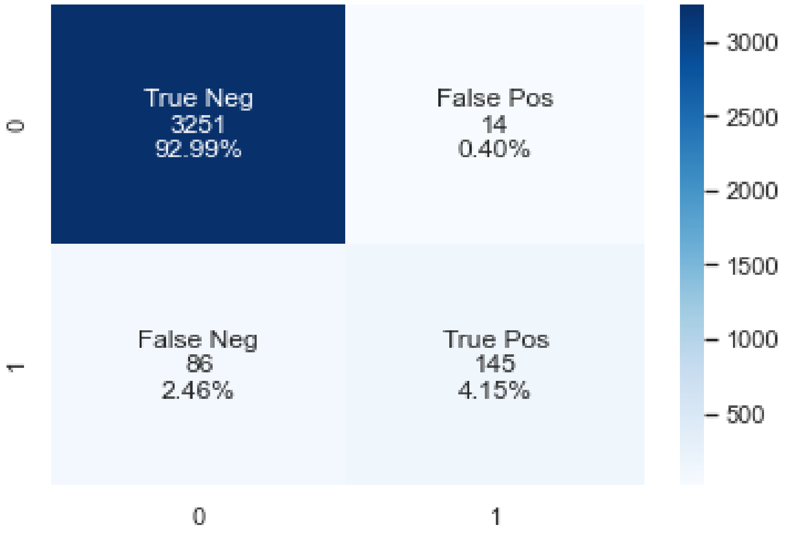


Based on the observations from the figure, it is apparent that the Mean Square Error of the logistic regression model is notably high, whereas the Mean Square Errors of both the SGD Classifier and Random Forest Classifier are considerably low. Consequently, for further analysis, it is advisable to employ either the SGD Classifier or the Random Forest Classifier due to their superior performance in minimizing Mean Square Error.

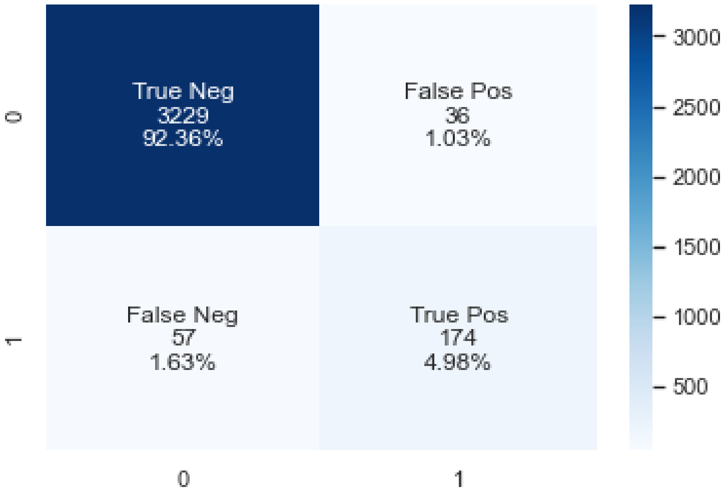
Confusion Matrices of Different Classifiers:

sA confusion matrix can be used to evaluate the quality of the project. The project aims to identify real and fake jobs.

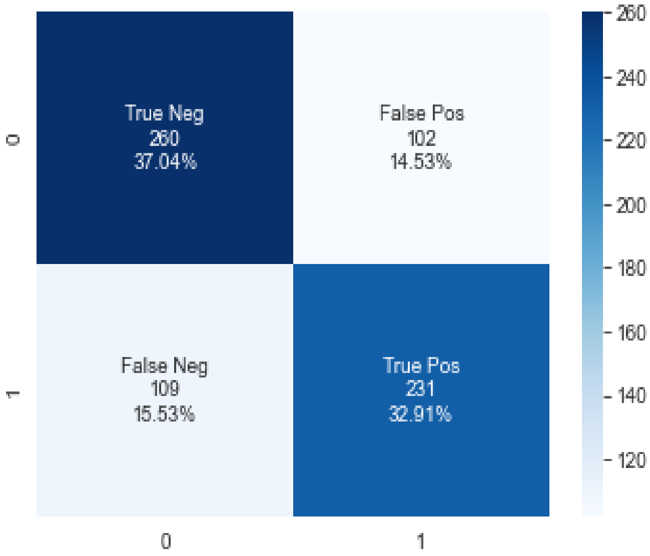
Multinomial Naïve Bayes Classifier



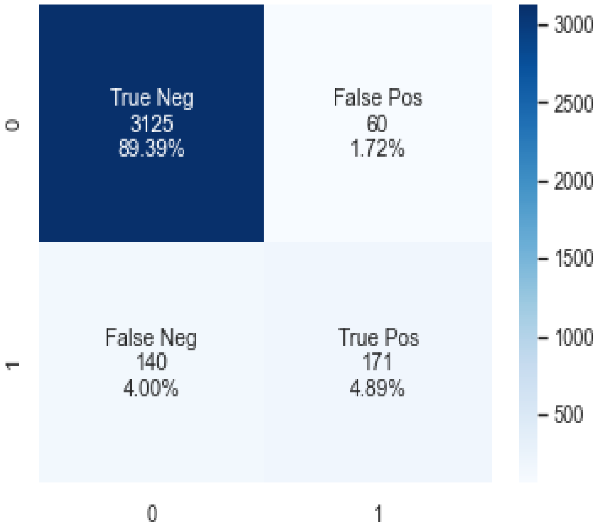
SGD Classifier



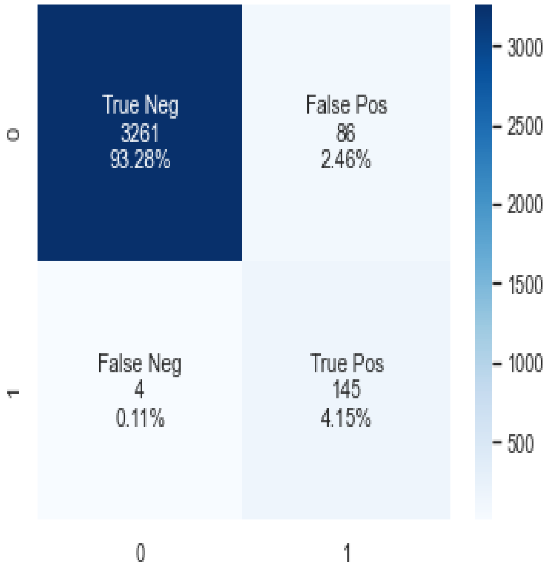
Logistic Regression Classifier



K-Nearest Neighborhood Classifier



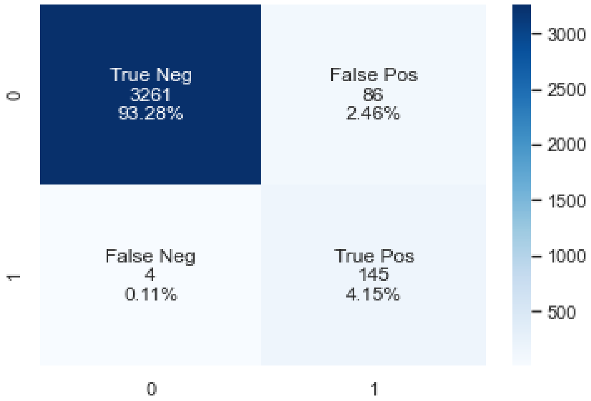
Random Forest Classifier



Final Model

RANDOM FOREST

The confusion matrix displays the following values –



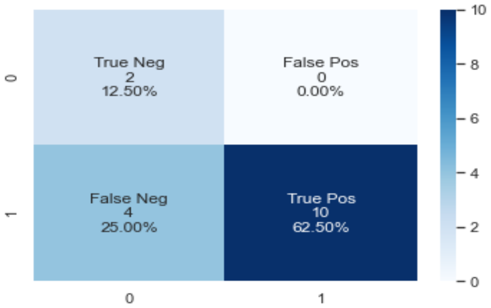
Categorized label, number of datapoints categorized under the label and percentage of data represented in each category. The test set has a total of 3347 real jobs and 149 fake jobs.

Performance Measure Metric	Random Forest Classifier
Accuracy	**97.48%**
F-1 Score	**77.28**
Cohen Kappa Score	**0.76**
MSE	**0.024**

Upon examining the confusion matrix, it becomes apparent that the model accurately identifies real jobs approximately 97.43% of the time, while fraudulent jobs are identified with slightly lower accuracy at

97.31%. However, there is a shortfall where the model misclassifies approximately 2.57% of the instances. This tendency for machine learning algorithms to favour dominant classes has been previously addressed.

Validation



- ➢ Accuracy score=81%
- ➢ F1 Score=80
- ➢ Cohen kappa Score= 0.3846153
- ➢ MSE=0.12

Conclusion

Employment scam detection plays a crucial role in safeguarding job-seekers by ensuring they receive legitimate offers from reputable companies. To combat this issue, various machine learning algorithms are proposed as countermeasures within this project's framework. Employing a supervised mechanism, several classifiers are utilized for employment scam detection, yielding notable findings:

1. Experimental results underscore the efficacy of the Random Forest classifier, which outperforms its counterparts. The proposed approach achieves an impressive accuracy of 97.48%, representing a significant improvement over existing methods.
2. A noteworthy aspect of the project is the identification of specific locations characterized by a high prevalence of fraudulent job postings. For instance, Bakersfield, California exhibits a concerning fake to real job ratio of 15:1, highlighting the need for enhanced monitoring in such areas.
3. An intriguing observation is the prevalence of fraudulent activity in entry-level job postings. It appears that scammers often target younger individuals with bachelor's degrees or high school diplomas seeking full-time employment opportunities. This underscores the importance of vigilance when navigating job offers in these categories.

s

Suggestion

Addressing the issue of fake job postings is a significant real-world challenge that demands proactive solutions. This project aims to contribute a potential remedy to this problem by meticulously pre-processing textual data and selecting relevant numerical fields to optimize results. By combining the outputs of multiple models, the objective is to attain the most effective outcomes while mitigating bias towards dominant classes.

One of the most demanding aspects of the project was the pre-processing of textual data, which required substantial effort due to its varied format and complexity. The dataset utilized in this endeavor exhibits a considerable imbalance, with the majority of jobs being real and only a few being fraudulent. Consequently, real jobs are being accurately identified to a significant extent. To address this imbalance, techniques like Synthetic Minority Over-sampling Technique (SMOTE) can be employed to generate synthetic samples for the minority class, thereby facilitating the creation of a more balanced dataset and potentially enhancing the overall results.

Reference

1. https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction
2. https://towardsdatascience.com/fake-job-predictor-a168a315d866
3. Bao, Y., Guan, Y., & Yan, S. (2020). Fake Job Posting Detection Based on Supervised Learning Algorithm. In 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 713-718). IEEE.
4. Yang, Z., & Huang, Z. (2019). Research on Classification and Detection of Fake Job Advertisements. In 2019 2nd International Conference on Computer Science and Advanced Materials Technologies (CSAMT) (pp. 493-497). IEEE.
5. Swigart, A., & Tovar, A. (2020). Detecting Fake Job Postings Using Natural Language Processing. arXiv preprint arXiv:2002.04869.
6. Kao, Y. C., Huang, J. C., & Hsu, W. L. (2020). Fake job advertisements detection using text mining techniques. In 2020 10th International Conference on Information Communication and Management (ICICM) (pp. 203-208). IEEE.
7. Fan, C., Chen, Y., & He, Y. (2020). A fake job detection method based on semantic analysis and ensemble learning. Journal of Intelligent & Fuzzy Systems, 39(1), 799-810.
8. Tong, Y., & Zhou, X. (2019). A Deep Learning-Based Approach to Detecting Fake Job Postings. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 307-317). Springer, Cham.
9. Sharma, P., & Joshi, A. (2020). Fake Job Detection: A Survey. In Advances in Cyber Security: Principles, Techniques, and Applications (pp. 399-412). Springer, Singapore.
10. Raj, S., & Kaul, R. (2019). Fake Job Detection Using Machine Learning Techniques. In 2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS) (pp. 1-6). IEEE.
11. Yang, S., Dong, H., & Wu, Y. (2020). A Hybrid Approach to Fake Job Posting Detection Using Machine Learning Algorithms. In International Conference on Network and System Security (pp. 287-299). Springer, Cham.
12. Wang, S., & Zeng, Y. (2020). Fake Job Detection Based on Convolutional Neural Network. In 2020 3rd International Conference on Computer Science and Software Engineering (CSASE) (pp. 263-268). IEEE.