Fake Job Recruitment Detection Using Machine Learning

Group- 69 CD
Authors:
VIJAY RAM CHALLAGUNDLA (AM.EN.U4CSE18269)
ACHYUTA VINDHYA (AM.EN.U4CSE18301)
T.D.CHARMILA (AM.EN.U4CSE18254)

GUIDE:
SUMESH KJ

Contents

1. Introduction
2. **Problem** Definition
3. Related Work
4. Proposed Method
5. Performance Evaluation/ Validation
6. Conclusion
7. References

Introduction:

Fake job news are presently the most frequently occurring problem in the world. This is because lack of knowledge about the news which are true or false.

If the person apply for the fake job the scammers will demand the money for the job. To detect whether the news is true or fake "Fake Job recruitment" Using machine learning was introduced.

At conceptual level, fake job recruitment has been classified into different types: the knowledge is then expanded to generalize machine learning(ML) models for multiple domains.

Problem Definition

The World Wide Web contains data in diverse formats such as documents, videos, audio, etc...The response that an article gets can be differentiated at a theoretical level to classify the article as real or fake.

The scammers provide users with a very lucrative job opportunity and later ask for money in return.

Background

- Employment scams are on the rise. According to CNBC ,the number of employment scams doubled in 2018 as compared to 2017.
- Economic stress and the impact of the corona virus have significantly reduced job availability and the loss of jobs form any individuals.
- Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. Most scammers do this to get personal information from the person they are scamming. Personal information can contain address, bank account details ,social security number etc.

Problem addressed/Your Work in brief

As our data set was imbalanced so it becoming a biased data set, so for overcoming that problem we used under sampling for majority outcomes.

Since the data provided has both numeric and text features models will be used on the text data and numeric data. The final output will be a combination of the two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not.

Related work

- **Spam detection**
- The problem of detecting non-genuine information sources through content-based analysis is thought to be solvable, at least in the domain of spam detection. Spam detection employs statistical machine learning techniques to determine whether text is spam or legitimate.
- Pre-processing of the text, feature extraction (i.e. bag of words), and feature selection based on which features lead to the best performance on a test dataset are all part of these techniques. After obtaining these features, they can be classified using Naive Bayes, Support Vector Machines, TF-IDF, or K-nearest Neighbor classifier All of these classifiers are characteristic of supervised machine learning, meaning that they require some labelled data in order to learn the function.

Related work

- **Benchmark Data set**
- It demonstrates previous work on fake news detection that is more directly
- related to our goal of using a text-only approach to make a classification. The authors not only create a new benchmark dataset of statements , but also show that significant improvements can be made in fine-grained fake news detection by using meta-data (i.e. Speaker, party, etc) to augment the information provided by the text.

Proposed Method

Block Diagram:

Data Scanning

Data Cleaning/Wrangling

Data Resizing

Training the ML Model

Testing the ML Model

Visualizing the data/result

Proposed Methods

- Naive Bayes
- Stochastic Gradient Descent
- Decision Tree Classifier
- Random Forest(Oversampling&undersampling)

Proposed Methods

- XGBoosting Classifier
- KNNeighbors Classifier
- GradientBoosting Classifier
- Support Vector Machine
- Logistic Regression

Module Description

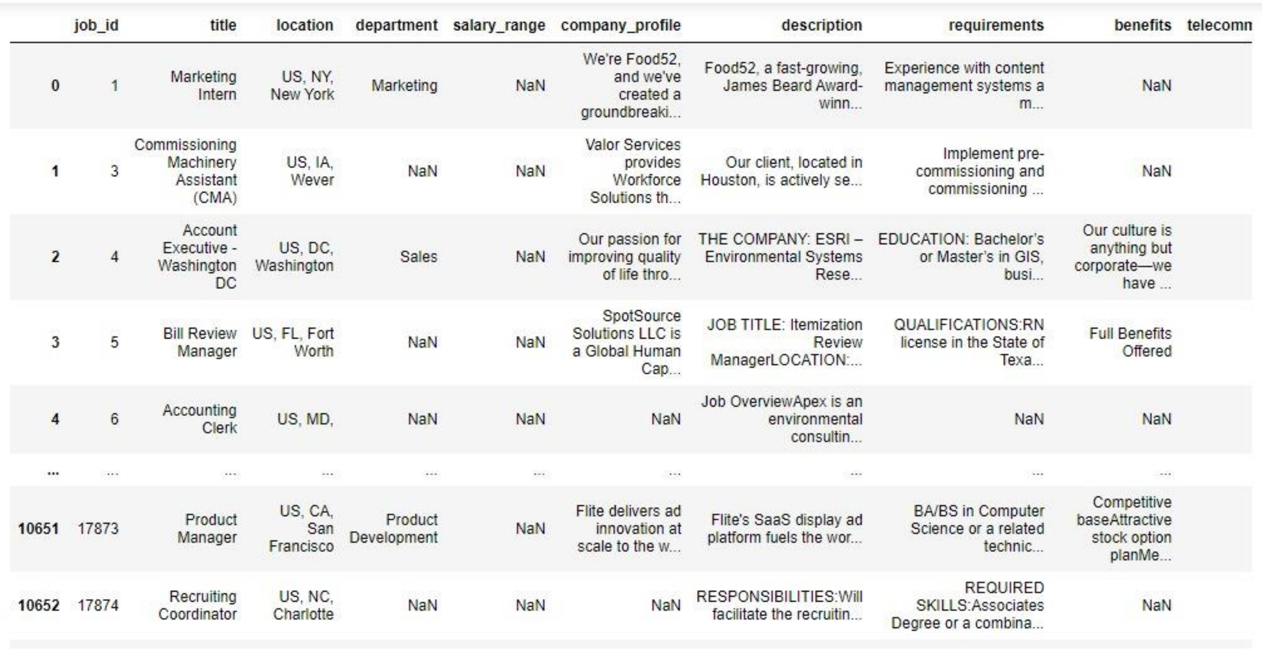
- The models will be evaluated based on two metrics:
- Accuracy: This metric is defined by this formula -

$$Accuracy = \frac{True\ Positve + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

- Recall-Score:
- The recall is the ratio `tp / (tp + fn)` where `tp` is the number of true positives and `fn` the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

Data set

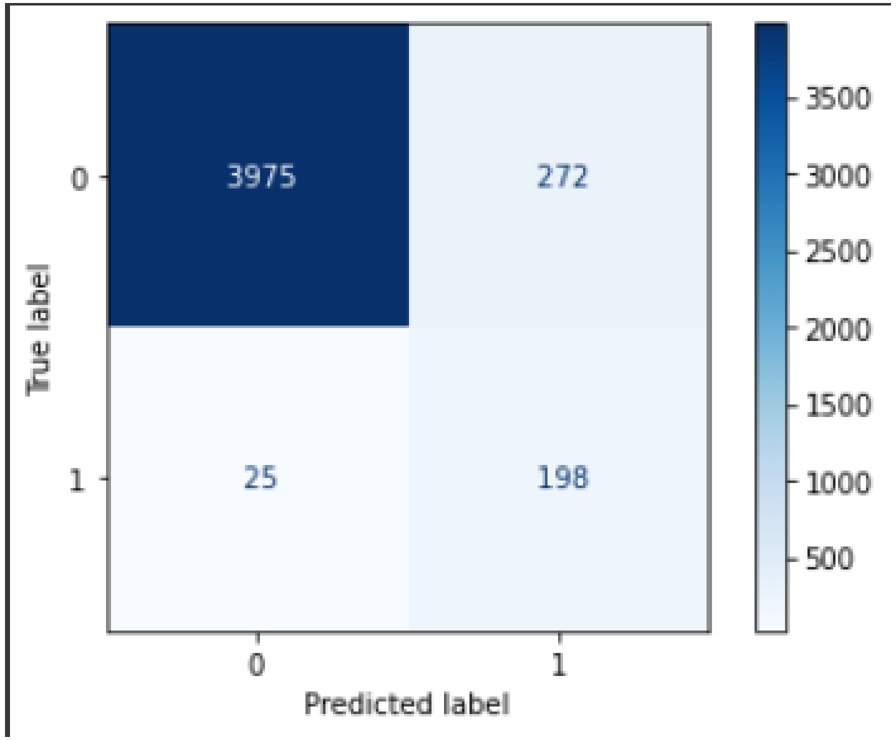
- The Data set we used is “fack_job_postings.csv”
- https://www.kaggle.com/vijayramchallagundla/fake-job-posting
- The dataset consists of 17,880 observations and 18 features.



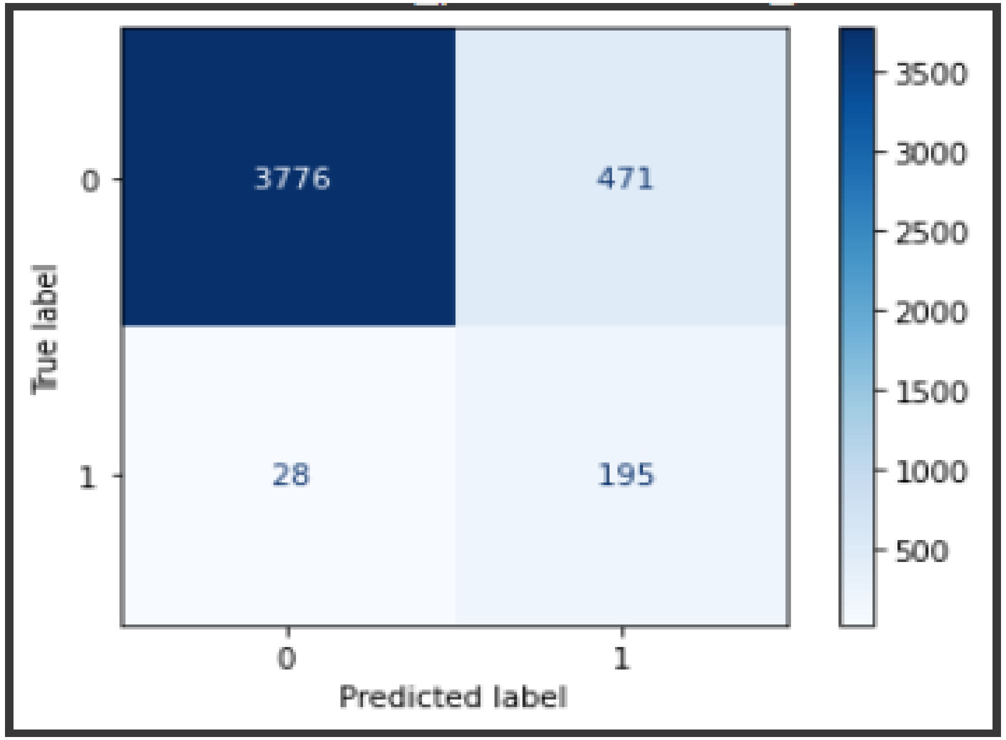
	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecomn
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	
1	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	
2	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 ...	
3	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	
4	6	Accounting Clerk	US, MD,	NaN	NaN	NaN	Job OverviewApex is an environmental consultin...	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	
10651	17873	Product Manager	US, CA, San Francisco	Product Development	NaN	Flite delivers ad innovation at scale to the w...	Flite's SaaS display ad platform fuels the wor...	BA/BS in Computer Science or a related technic...	Competitive baseAttractive stock option planMe...	
10652	17874	Recruiting Coordinator	US, NC, Charlotte	NaN	NaN	NaN	RESPONSIBILITIES:Will facilitate the recruitin...	REQUIRED SKILLS:Associates Degree or a combina...	NaN	

Confusion Matrix for Algorithm

- Multi Naive Bayes Classifier

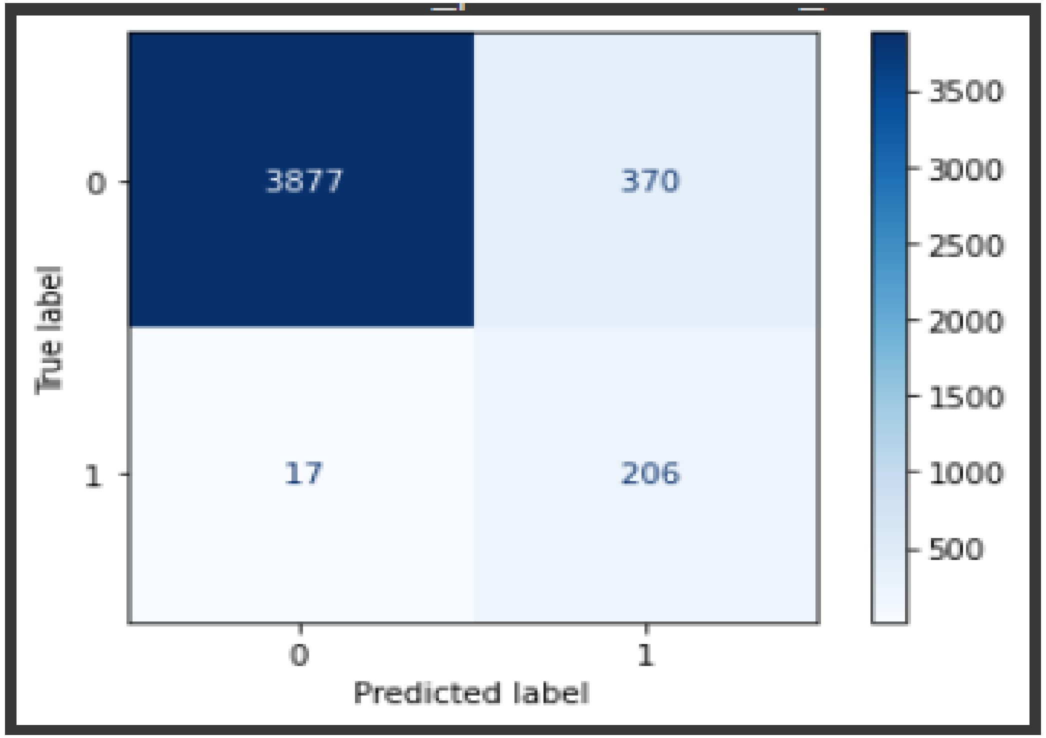


- Decision Tree Classifier

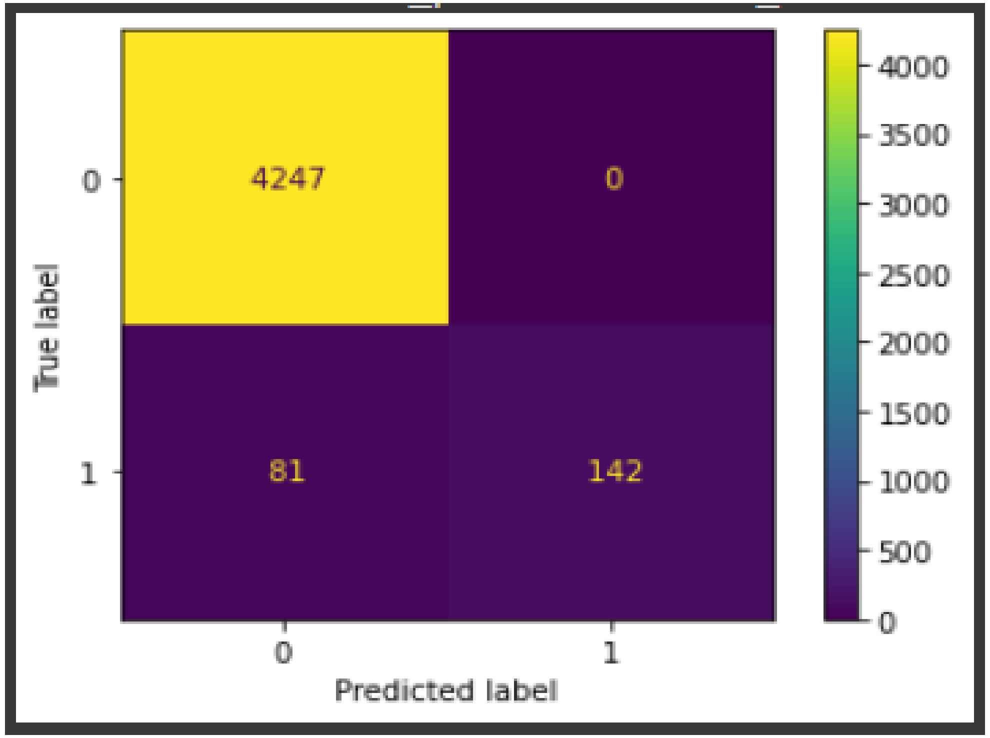


Confusion Matrix for Algorithm

- SGD Classifier

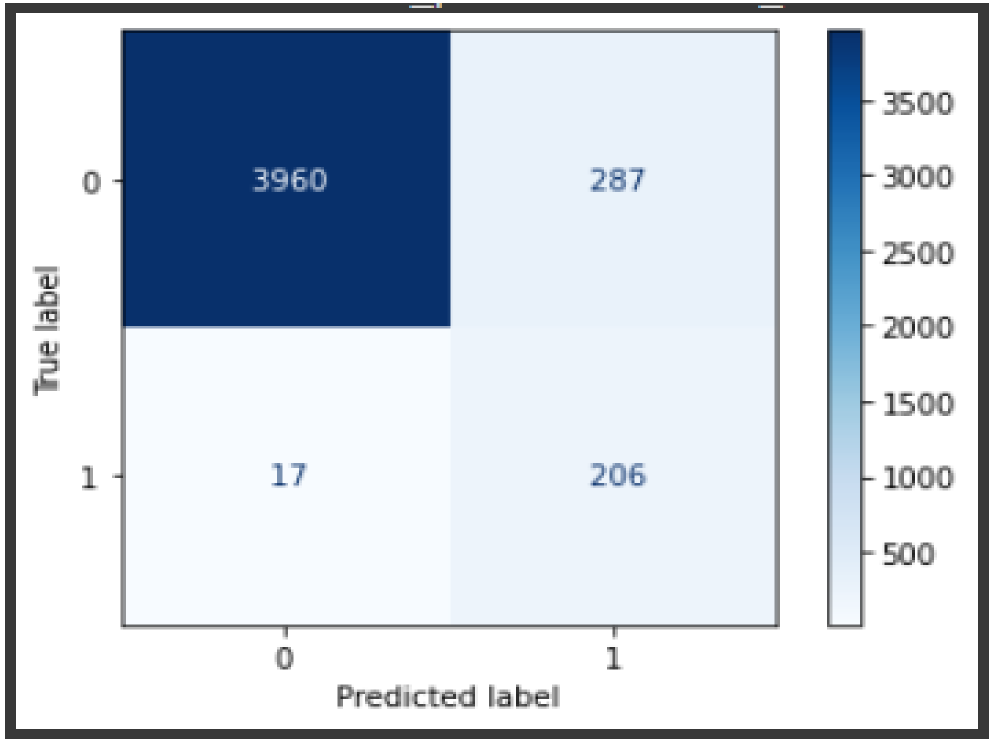


- Random Forest Classifier

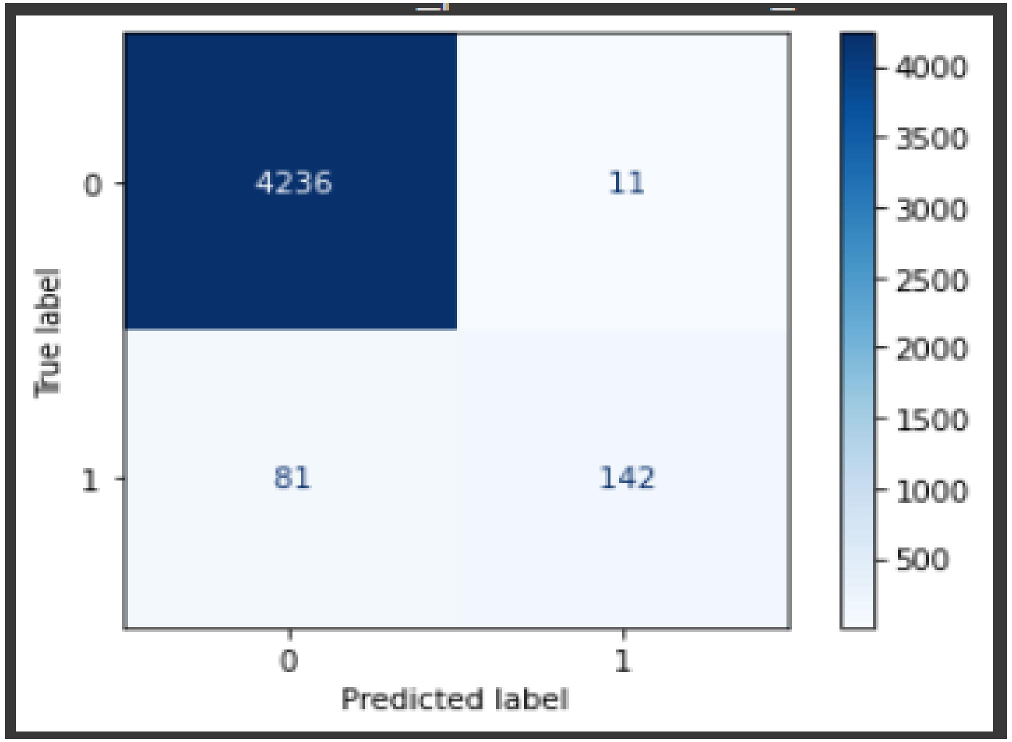


Confusion Matrix for Algorithm

- Random Forest Classifier Under-Sampling

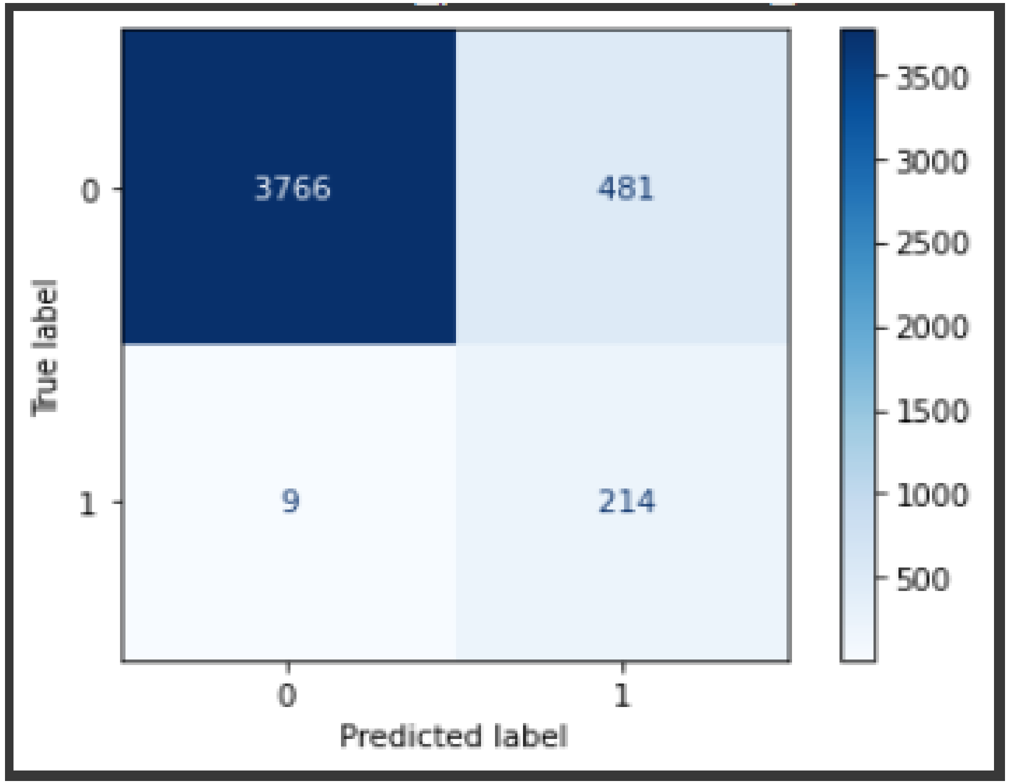


- Random Forest Classifier Over-Sampling

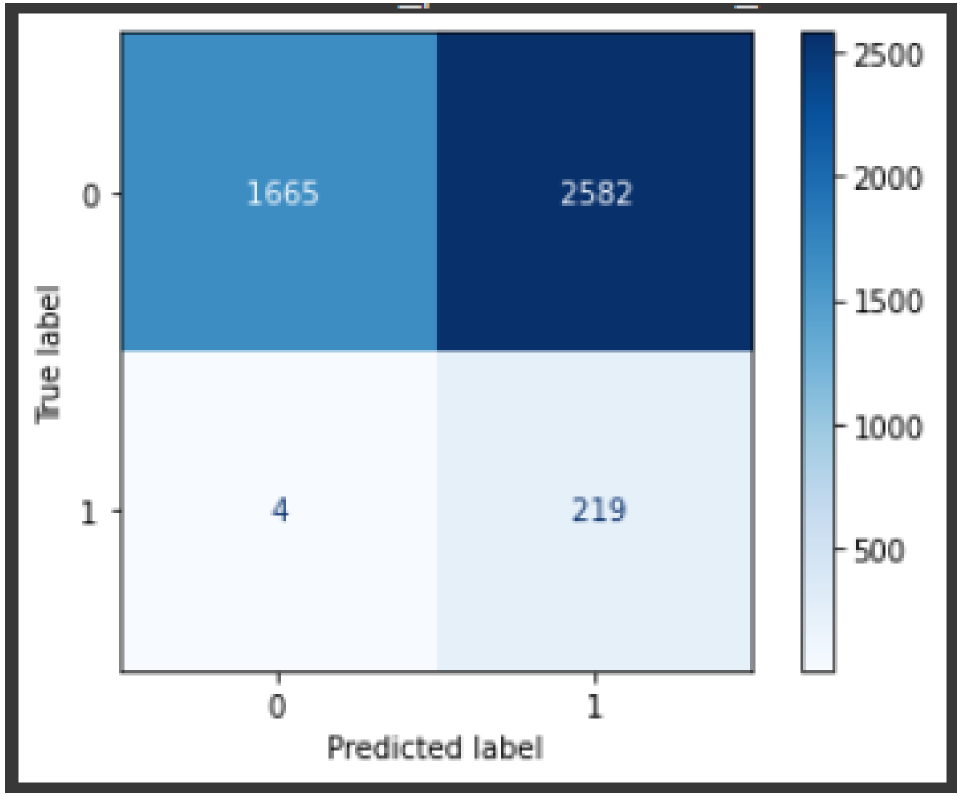


Confusion Matrix for Algorithm

- XGB Classifier

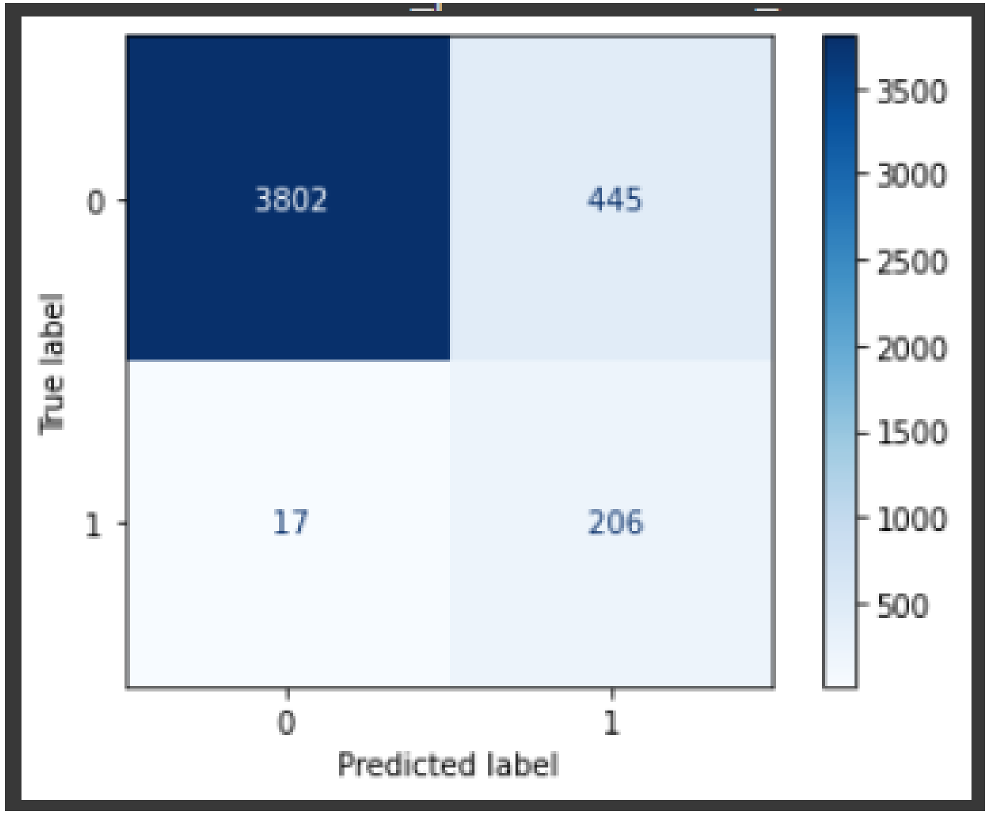


- K Nearest Neighbors Classifier

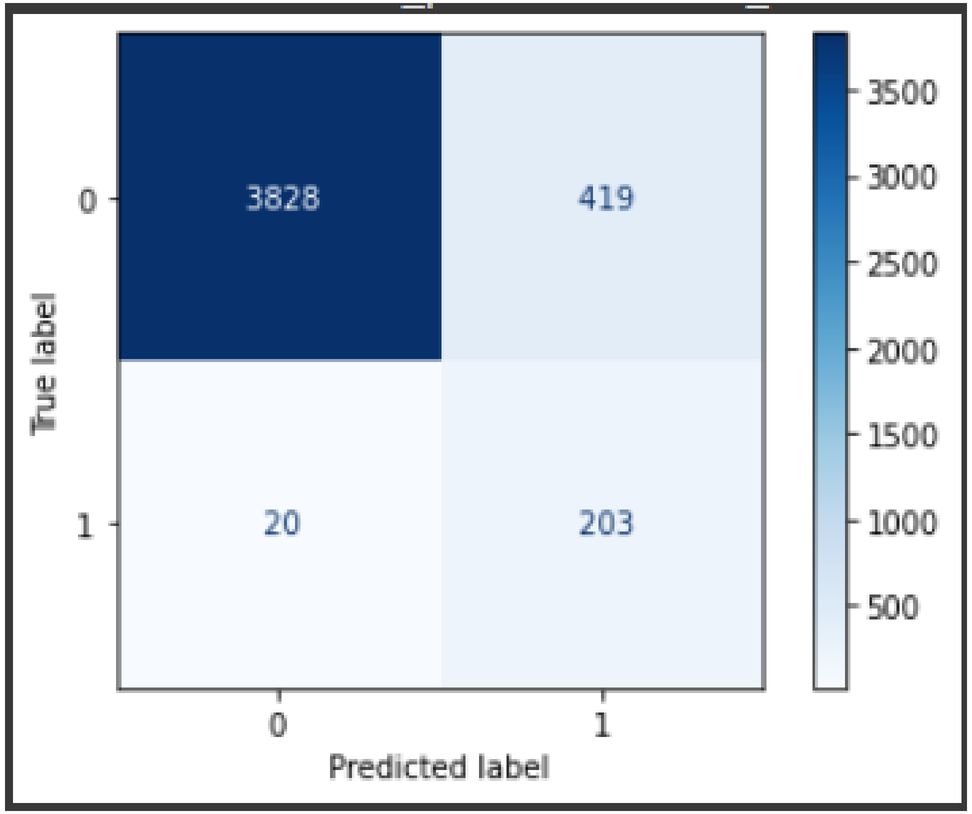


Confusion Matrix for Algorithm

- Gradient Boosting Classifier

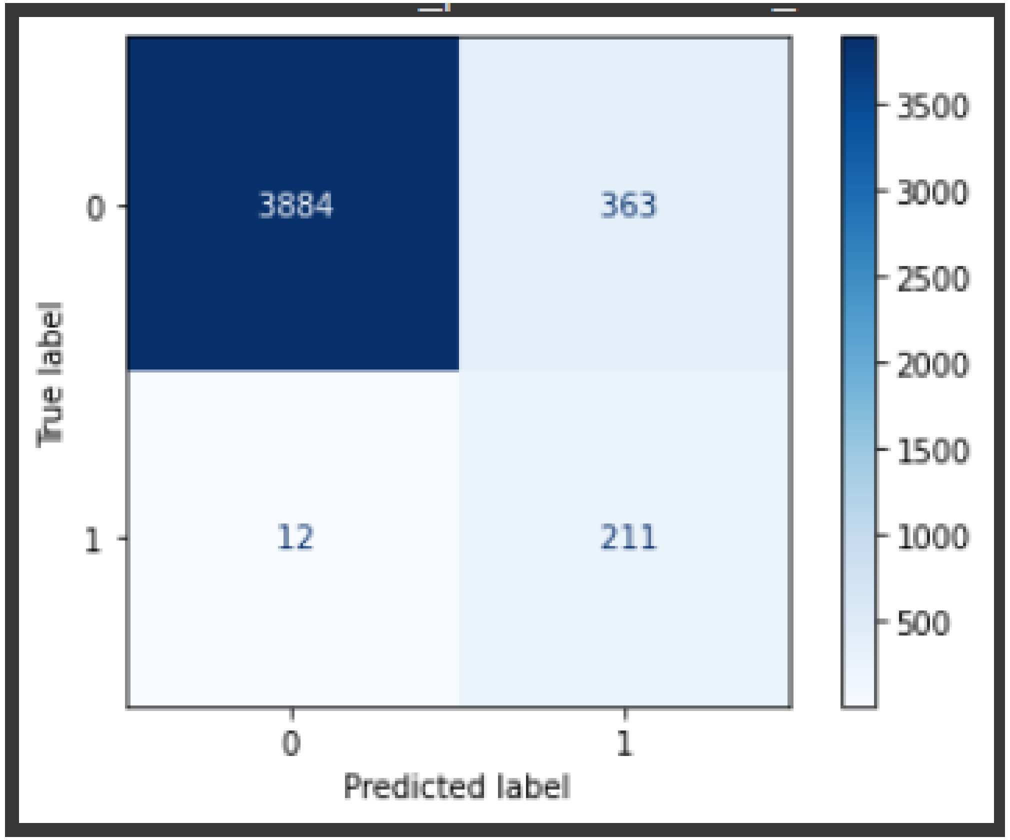


- Support Vector Machine Classifier



Confusion Matrix for Algorithm

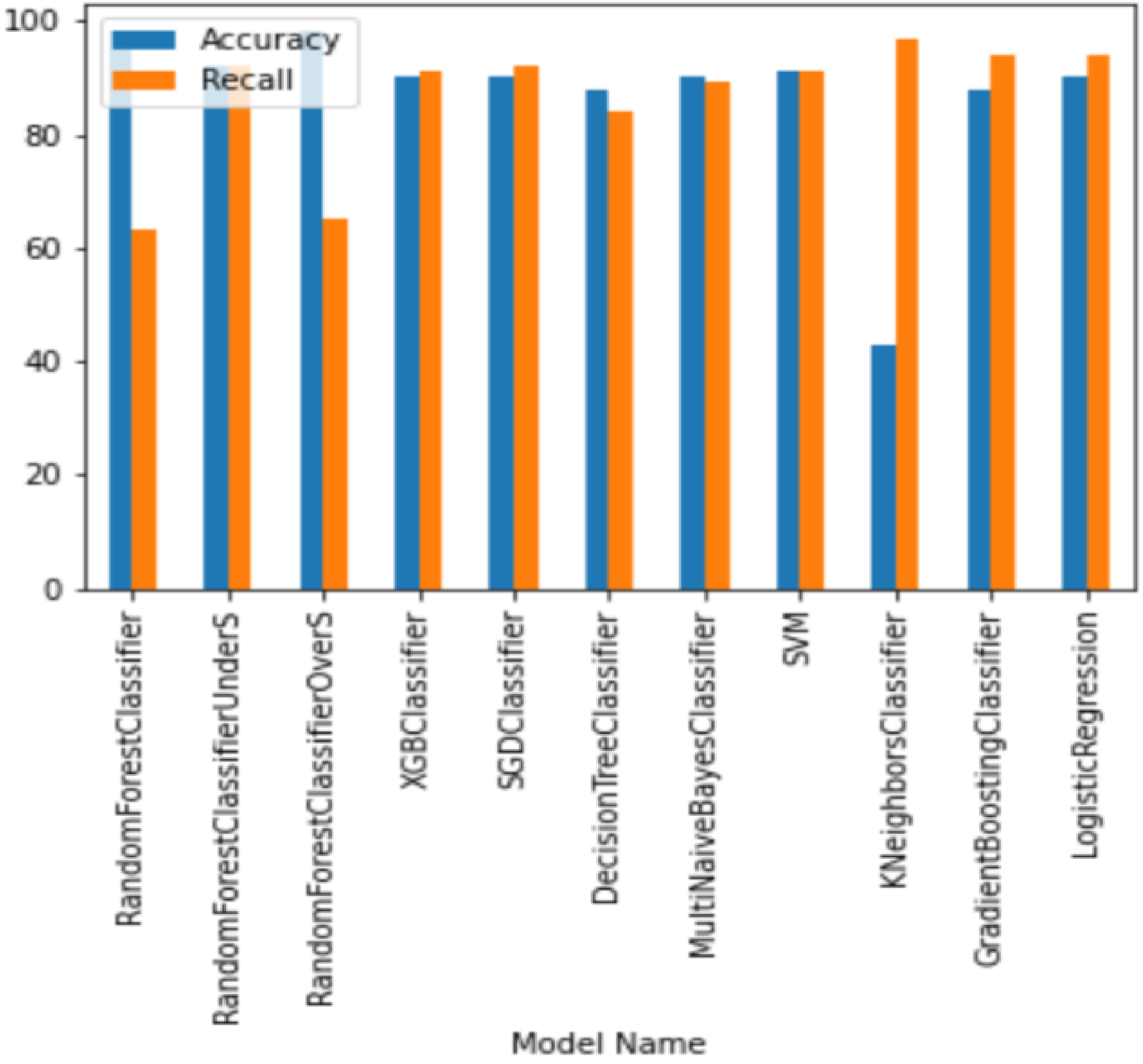
- Logistic Regression Classifier



Results

S.No	Algorithm	Accuracy	Recall Score
1	Multi Naive Bayes Classifier	0.90648	0.89686
2	Decision Tree Classifier	0.88389	0.84304
3	SGD Classifier	0.90425	0.92376
4	Random Forest Classifier	0.98187	0.63677
5	Random Forest Classifier Under-Sampling	0.92930	0.92376
6	Random Forest Classifier Over-Sampling	0.98031	0.65022
7	XGB Classifier	0.90357	0.91479
8	K Nearest Neighbors Classifier	0.43467	0.97309
9	Gradient Boosting Classifier	0.88926	0.94170
10	Support Vector Machine Classifier	0.91073	0.91928
11	Logistic Regression Classifier	0.90156	0.94170

Performance Measures



Conclusion

- The dataset that is used in this project is very unbalanced. Most jobs are real, and few are fraudulent. Due to this, real jobs are being identified quite well. Certain techniques like under sampling, over sampling, SMOTE used to generate synthetic minority class samples. So the balanced dataset has performed to generate better results.
- We perform all algorithms for our data set. We find accuracy and recall for them
- We got similar values for all algorithms ,and we compared all the algorithms and choose the best one.

Further work

- In Further we would like to improve our work to make this this project look some attractive by connecting this data to one WEB APPLICATOIN.

- We would like to make this as user friendly, and we got to conclude that making a chrome extension would be better.

References

- [1] Fake News Detection Using Machine Learning Vijaya Bal-pande Kasturi Baswe Kajol Somaiya Achal Dhande Prajwal Mirehttps://doi.org/10.32628/CSEIT12173115
- [2] Fake News Detection Using Machine Learning Ensemble Methods.Iftikhar Ahmad ,Muhammad Yousaf, Suhail Yousaf , and Muhammad OvaisAhmad , https://doi.org/10.1155/2020/8885861
- [3] ftikharAhmad,1MuhammadYousaf,1SuhailYousaf,1andMuhammad Ovais Ahmad Volume 2020 —Article ID 8885861 —https://doi.org/10.1155/2020/8885861
- [4] . Smitha and R. Bharath, "Performance Comparison of Machine LearningClassifiers for Fake News Detection," 2020 Second International Conferenceon Inventive Research in Computing Applications (ICIRCA), 2020, pp. 696-700, doi: 10.1109/ICIRCA48905.2020.9183072.
- [5] n Intelligent Model for Online Recruitment Fraud Detection Bandar Al-ghamdi, Fahad AlharbyNaif Arab University (NAUSS), Riyadh, KSA.DOI:10.4236/jis.2019.103009

References

- [6] . I. Manzoor, J. Singla and Nikita, ''Fake News Detection Using Machine Learning approaches: A systematic Review,'' 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 230-234,doi: 10.1109/ICOEI.2019.8862770.
- [7] Fake News Detection Using Machine Learning Approaches To cite this article: Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040
- [8] Fake News Detection using Machine Learning Ensemble Methods Volume2020—Article ID 8885861 — https://doi.org/10.1155/2020/8885861
- [9] amirBandyopadhyaApril2020DOI:10.14445/22315381/IJETT-V68I4P209S
- [10] Fake News Detection DOI: 10.1109/SCEECS.2018.8546944

References

- [11]Fake News Detection Using Machine Learning Approaches To cite thisarticle: Z Khanam etal 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040
- [12] Automating the Detection of Cyberstalking
- [13] Fake News Detection DOI: 10.1109/SCEECS.2018.8546944
- [14] Detection of SPAM Attacks in the Remote Triggered WSN Experiments
- [15]K. R. Vidya Kumari & C. R. Kavitha (2018). Spam Detection Using Machine Learning in R (International Conference on Computer Networks and Communication Technologies pp 55–64)
- [16]N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113,

doi: 10.1109/ICIRCA48905.2020.9183098.

THANK YOU