# Finally, Report

# Fake Job Predictor

## Abstract

Employment scams are on the rise. According to CNBC, the number of employment scams doubled in 2018 as compared to 2017. The current market situation has led to high unemployment. Economic stress and the coronavirus's impact have significantly reduced job availability and job loss for many individuals. A case like this presents an appropriate opportunity for scammers. 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 addresses, bank account details, social security numbers, etc. The scammers provide users with a very lucrative job opportunity and later ask for money in return. Or they require investment from the job seeker with the promise of a job. This is a dangerous problem that can be addressed through Machine Learning techniques and Natural Language Processing (NLP).

## Problem Statement

This project aims to create a classifier that will have the capability to identify fake and real jobs. The final result will be evaluated based on two different models. Since the data provided has both numeric and text features one model will be used on the text data and the other on numeric data. The final output will be a combination of the two.

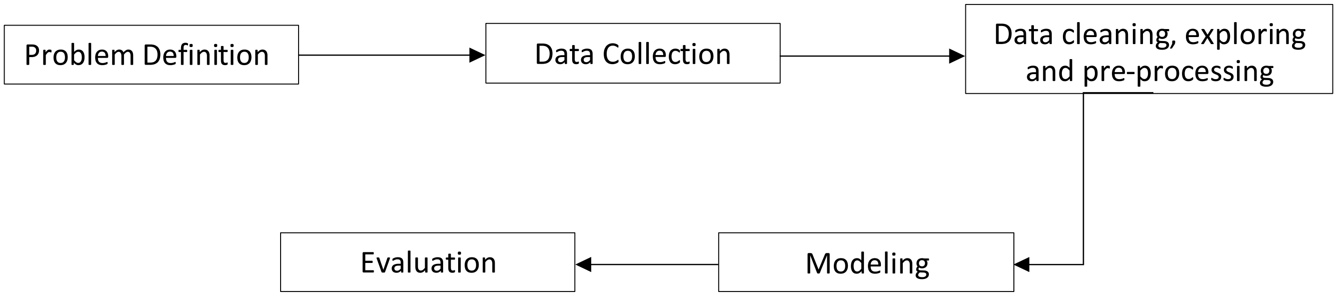
The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not.

## Design

This project uses data provided from Kaggle. This data contains features that define a job posting. These job postings are categorized as either real or fake. Fake job postings are a very small fraction of this dataset. That is as excepted. We do not expect a lot of fake jobs postings. This project follows five stages.

The five stages adopted for this project are:

1. Problem Definition (Project Overview, Project statement and Metrics)
2. Data Collection
3. Data cleaning, exploring and pre-processing
4. Modeling
5. Evaluation



## Data

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

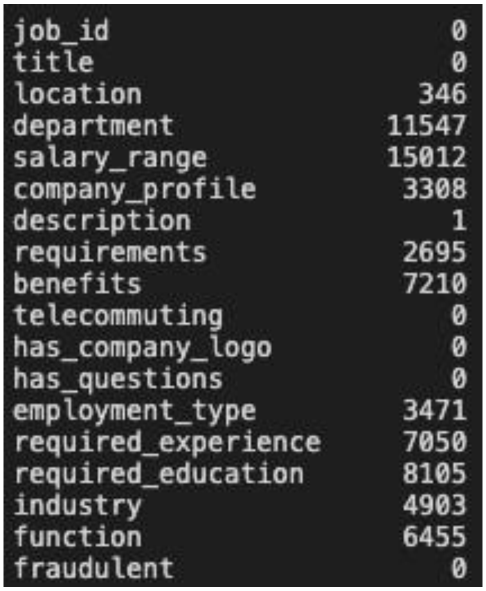
A brief definition of the columns is given below:

#Cl : it's mean Colum

| Cl n | Cl title | Cl Description |
|---|---|---|
| 1 | job_id | Identification number given to each job posting |
| 2 | title | A name that describes the position or job |
| 3 | location | Information about where the job is located |
| 4 | department | Information about the department this job is offered by |
| 5 | salary_range | Expected salary range |
| 6 | company_profile | Information about the company |
| 7 | description | A brief description about the position offered |
| 8 | requirements | Pre-requisites to qualify for the job |
| 9 | benefits | Benefits provided by the job |
| 10 | telecommuting | Is work from home or remote work allowed |
| 11 | has_company_logo | Does the job posting have a company logo |
| 12 | has_questions | Does the job posting have any questions |
| 13 | employment_type | 5 categories – Full-time, part-time, contract, temporary and other |
| 14 | required_experience | Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable |
| 15 | required_education | Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college |

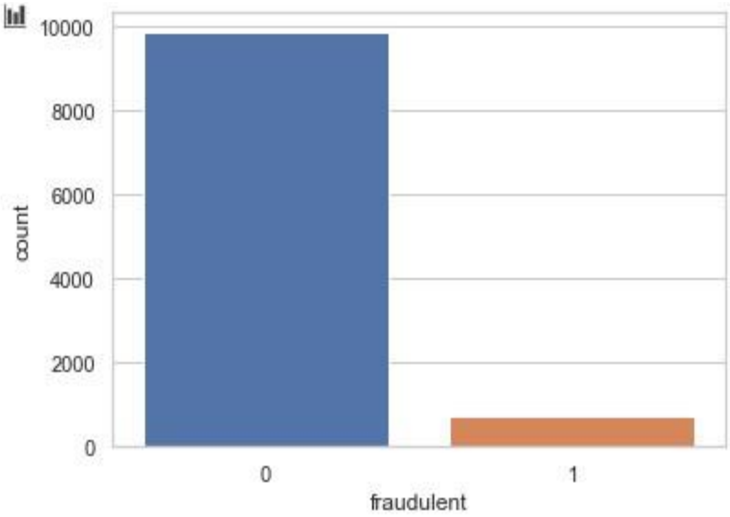
| | | |
|---|---|---|
| | | coursework, professional, some high school coursework, vocational |
| 16 | Industry | The industry the job posting is relevant to |
| 17 | Function | The umbrella term to determining a job's functionality |
| 18 | Fraudulent | The target variable ◊ 0: Real, 1: Fake |

Since most of the datatypes are either Booleans or text a summary statistic is not needed here. The only integer is job_id which is not relevant for this analysis. The dataset is further explored to identify null values.



```
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
```

The dataset is highly unbalanced with 9868 (93% of the jobs) being real and only 725 or 7% of the jobs being fraudulent. A count plot of the same can show the disparity very clearly.



## Algorithms

Based on the initial analysis, it is evident that both text and numeric data is to be used for final modeling. Before data modeling a final dataset is determined. This project will use a dataset with these features for the final analysis:

1. telecommuting
2. fraudulent
3. ratio: fake to real job ratio based on location
4. text: combination of title, location, company profile, description, requirements, benefits, required experience, required education, industry and function
5. character count: Count of words in the textual data Word count histogram

Further pre-processing is required before textual data is used for any data modeling.

The algorithms and techniques used in project are:

1. Natural Language Processing
2. Naïve Bayes Algorithm
3. SGD Classifier

Naïve Bayes and SGD Classifier are compared on accuracy and F1-scores and a final model is chosen. Naïve Bayes is the baseline model, and it is used because it can compute the conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event, encoding those probabilities is extremely useful.

A comparative model, SGD Classifier is used since it implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification. This classifier will need high penalties when classified incorrectly. These models are used on both the text and numeric data separately and the final results are combined.

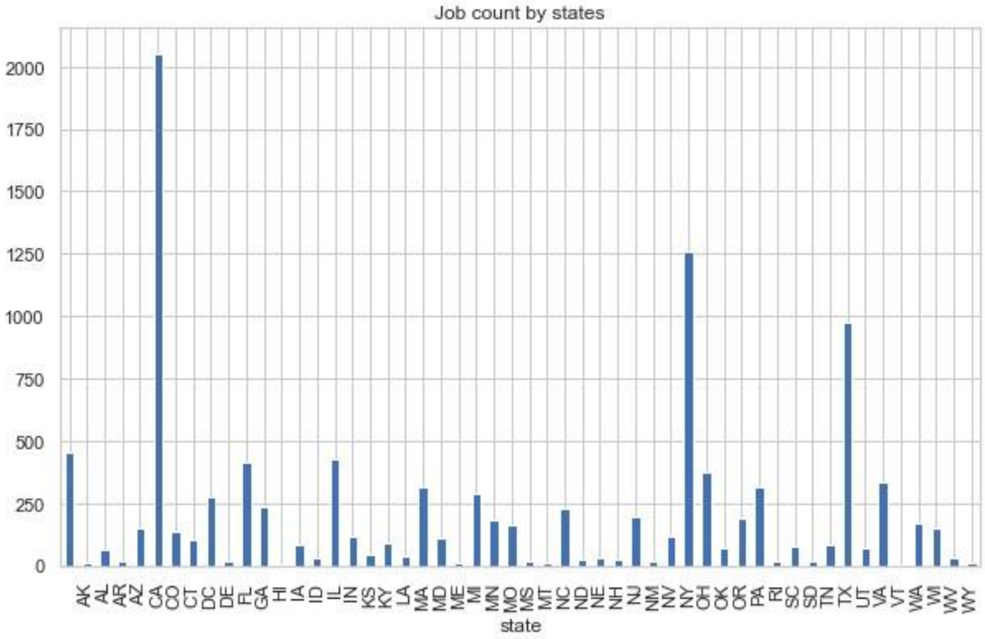
## Benchmark

The benchmark model for this project is Naïve Bayes. The overall accuracy of this model is 0.971 and the F1-score is 0.744. The reason behind using this model has been elaborated above. Any other model's capabilities will be compared to the results of Naïve Bayes.

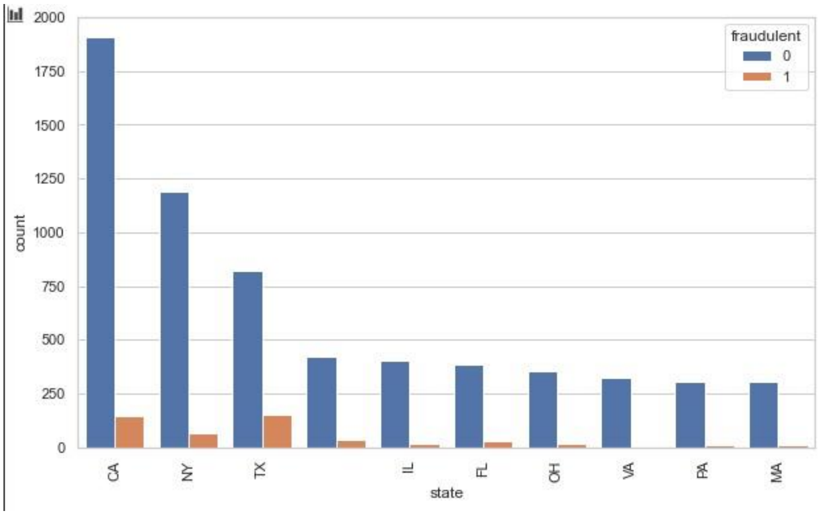
## Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Tableau for interactive visualizations

## Communication



The graph above shows which states produces the greatest number of jobs. California, New York and Texas have the highest number of job postings. To explore this further another bar plot is created. This barplot shows the distribution of fake and real jobs in the top 10 states.

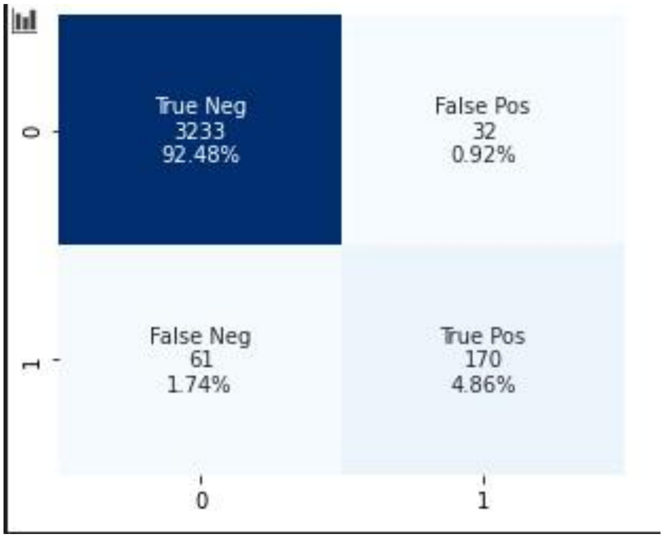


The graph above shows that Texas and California have a higher possibility of fake jobs as compared to other states. To dig one level deeper into and include states as well a ratio is created. This is a fake to real job ratio based on states and citifies.

## Conclusion

Free-Form Visualization

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



The confusion matrix above displays the following values – categorized label, number of data points categorized under the label and percentage of data represented in each category. The test set has a total of 3265 real jobs and 231 fake jobs. Based on the confusion matrix it is evident that the model identifies real jobs 99.01% of the times. However, fraudulent jobs are identified only 73.5% of the times. Only 2% of the times has the model not identified the class correctly. This shortcoming has been discussed earlier as well as Machine Learning algorithms tend to prefer the dominant classes.

## Reflection

Fake job postings are an important real-world challenge that require active solutions. This project aims to provide a potential solution to this problem. The textual data is pre-processed to generate optimal results and relevant numerical fields are choosing as well. The output of Multiple models is combined to produce the best possible results. This is done to reduce the bias that a machine learning model has towards the dominant class.

## Improvement

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 SMOTE can be used to generate synthetic minority class samples. A balanced dataset should be able to generate better results.