Fake Job Prediction

Milestone: Data Collection and Processing

Group: 15

Meenal Lnu

Pulkit Manchanda

(848) 667-4233

(857) 328-2523

lnu.me@northeastern.edu

manchanda.pu@northeastern.edu

Percentage of Contribution by Meenal: 50%

Percentage of Contribution by Pulkit: 50%

Signature of Meenal: *Meenal Lnu*

Signature of Pulkit: *Pulkit Manchanda*

Submission Date: February 28, 2022

Problem Setting

We are now living in a world where almost everything is done online. From learning to shopping, we do everything from our mobiles, tablets, laptops and computers. Like every other need of theirs, even for jobs, people search online. Multiple websites provide thousands of jobs listings each day.

World is right now affected by coronavirus. Millions of people are losing their jobs, unemployment has become a serious issue and people are desperately in search of jobs. This gives online scammers an opportunity to take advantage of such desperation. Many fake jobs are posted on the internet every single day. These fake postings look quite genuine. They may also have a recruitment process, a company website, almost like how a genuine job posting would be.

But, if we have a closer look, we can differentiate between fake and genuine postings. Most fake postings will not have a company logo or telephone number. The email replies might come from a personal email address. The biggest red flag would be, the company asking for credit cards details, or requesting to transfer some funds. All these situations are suspicious that something is not correct.

Problem Definition

This project aims to create a classifier that will have the capability to identify fake and real jobs. The model will take in all relevant job posting data and produce a final result determining whether the job is real or not.

There is a constant search for a job, and we see a daily rise in fake job postings. Thus, the main goal is to distinguish fake jobs from real ones by means of classification Additionally we hope to help with these through our project:

- Candidates in identification of genuine job postings
- Recruitment Counsellors to suggest real job openings to clients
- Job Portals to decide whether to consider the job openings to be displayed on their webpage
- Police in identifying employment scams

Data Sources

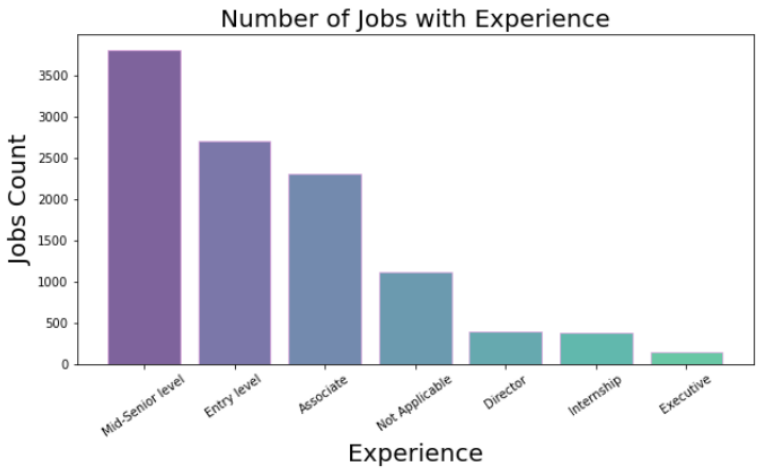
We have extracted the Dataset from Kaggle https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction which had been uploaded from the University of the Aegean - Laboratory of Information & Communication Systems Security http://emscad.samos.aegean.gr/

Data Description

Our dataset, The Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset. It contains 17,014 legitimate and 866 fraudulent job ads published between 2012 to 2014. There are a total of 17 features contributing to an online job posting. These are broadly divided into following categories:

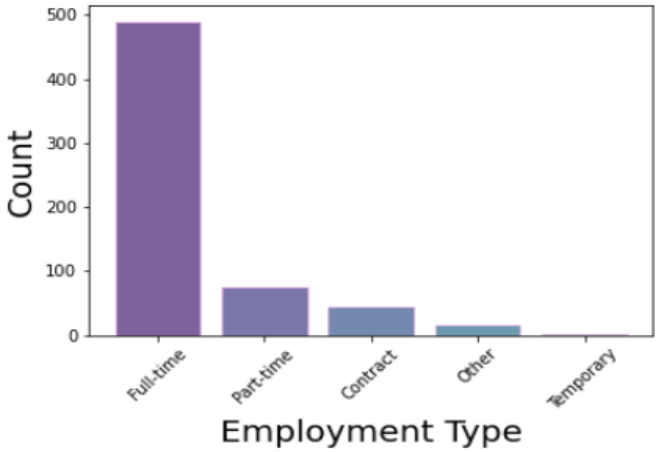
1. **Basic features** - The features which give the basic information of a job opening. This includes Title, Department, Location, Salary Range.
2. **Descriptive features** - The features which give an insight into the company and job details. This includes company profile, description, requirements, benefits.
3. **Binary features** - Whether the company has logo, questions, telecommuting
4. **Nominal Variables** - Employment type, required experience, required education, industry, function

Data Exploratory Analysis



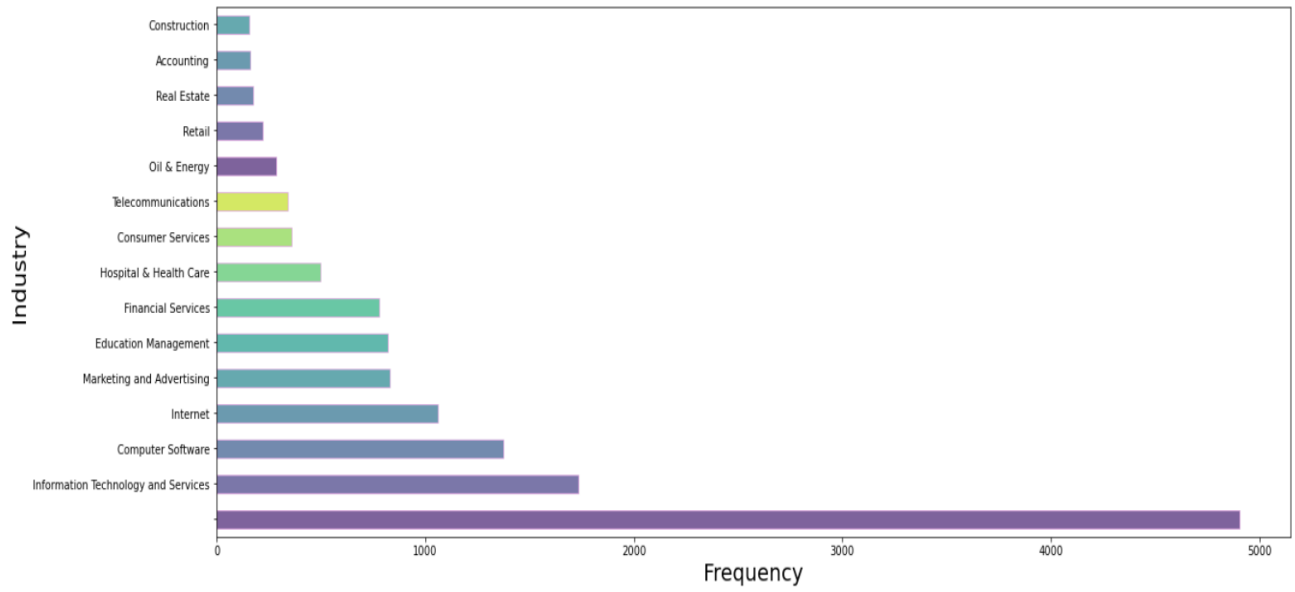
Required level of Experience vs count of the fraudulent jobs posted

As seen from the graph, Mid senior and Entry level jobs are generally more susceptible to fraud than other experience level jobs



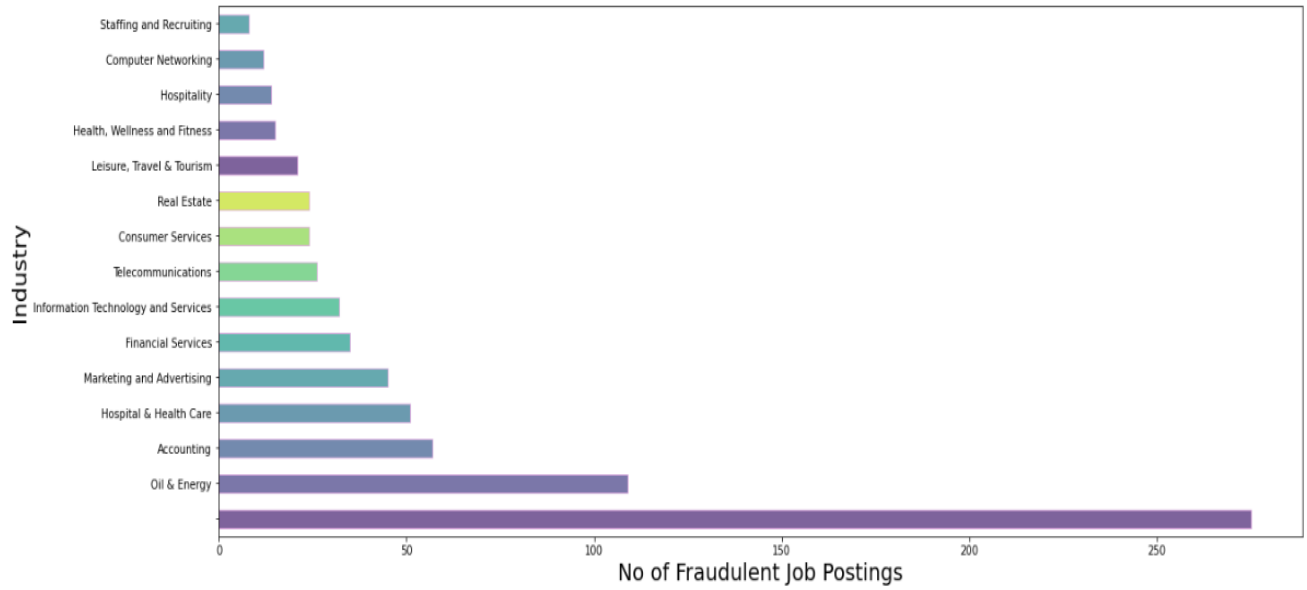
Employment type vs count of the fraudulent jobs posted

As seen from the graph, Full-time jobs are clearly more susceptible to fraud than Part-time or Contract-Based jobs.



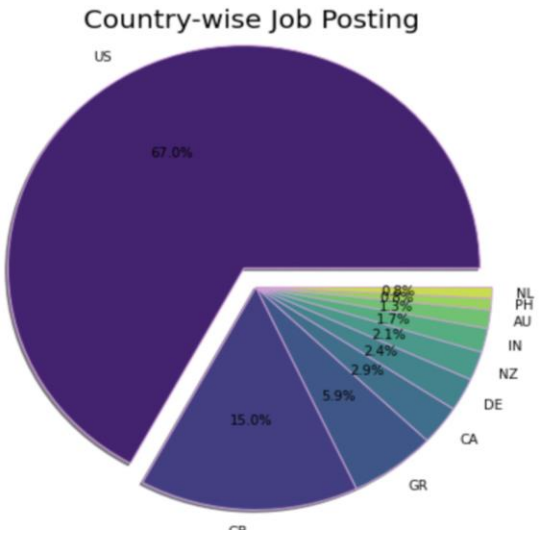
Industry with Max Job posting counts

We can observe that Tech related Industries like Information Technology & Services, Computer Science, Internet make up a large portion of the Job postings



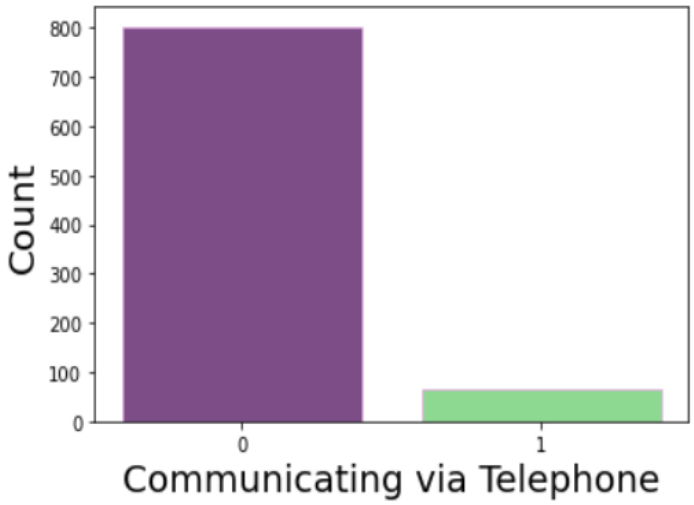
Industry with Max Fake job postings

We can observe that the high-paying, high-risk jobs in the industries like Oil & Energy, Accounting, Health Care, Advertising, Finance seem to attract scammers more and have a higher number



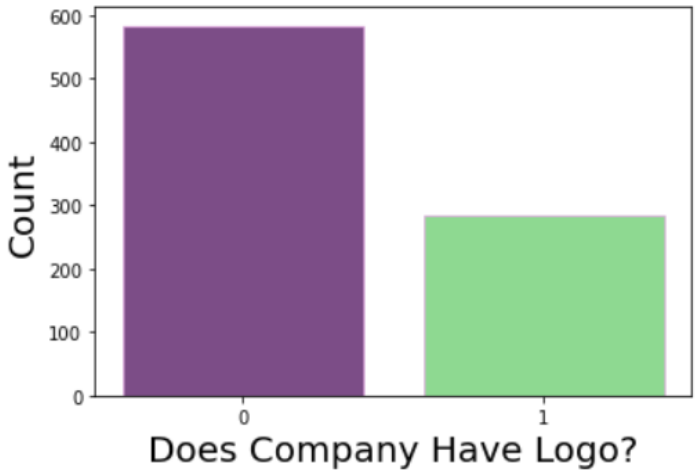
Country-wise Number of Job Posting

Almost 70% of the Companies Job Openings Uploaded on Employment Portals are from US based Companies



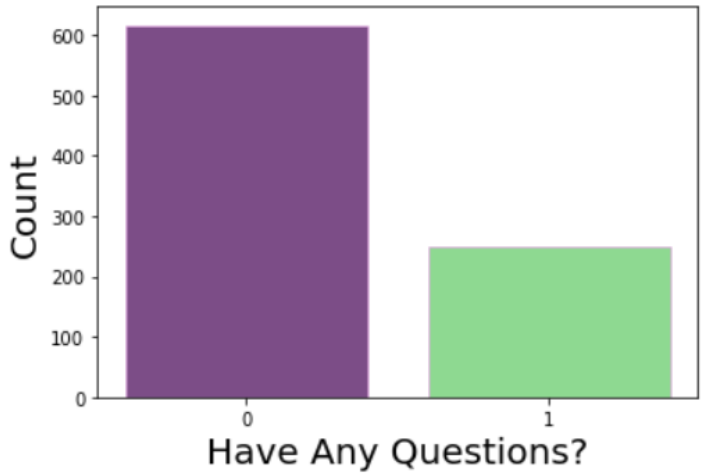
Job Openings with Contact Information like Contact Number

Here, we can depict that most of the companies has mentioned their Contact Numbers along with their job Openings.



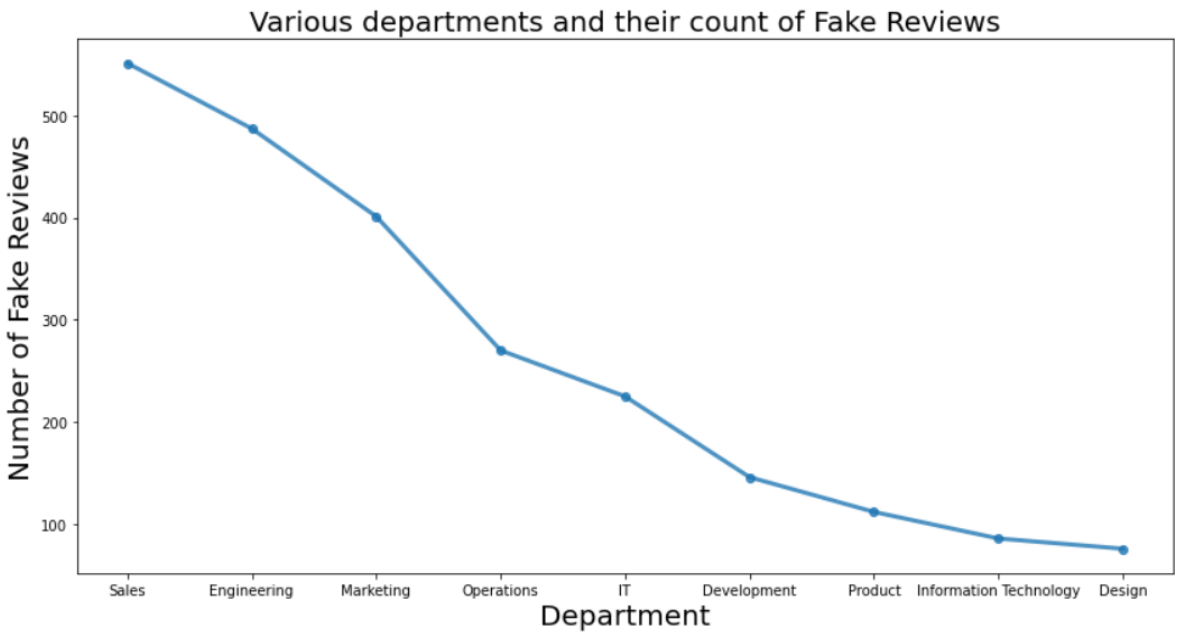
Job Openings with their Official Trademark

The Frequency of the Companies which have their Official Trademark while posting Job Openings on Job Portals



Job Openings with Query Section

Almost 30% of the Job Openings don't mention the Query Section related to respective Job Openings.



Various Departments and count of their Fake Reviews

English Teacher Abroad	311
Customer Service Associate	146
Graduates: English Teacher Abroad (Conversational)	144
English Teacher Abroad	95
Software Engineer	86
English Teacher Abroad (Conversational)	83
Customer Service Associate - Part Time	76
Account Manager	75
Web Developer	66
Project Manager	62

Most Frequent Roles for which Job Posting were made

Customer Service Associate	145
Software Engineer	71
Account Manager	64
Web Developer	50
Account Executive	40
Customer Service Associate	40
Customer Service Team Lead	39
Product Manager	39
Senior Software Engineer	37
Sales Representative	37

Job title with max full-time opportunities

Fraudulent jobs WordCloud

Actual jobs WordCloud

From the two WordClouds, we can see some differences. Fake jobs descriptions focus on words like data entry, equipment, design, technical, product etc. which are different from the words used to describe real job postings.

Data Pre Processing (prper heading and elaborate mmre with img)

From the data, we have filtered out all the fake jobs and performed exploratory data analysis to get more insights about such fake postings. The problems with the dataset that we found, and ways used to resolve them are as follows:

- **Over Sampling** - We observed that the target variable is imbalanced and skewed towards the 'Not fraudulent' label. To solve this problem, we considered Sampling technique which balances the records from both the classes and found **Random Over Sampler** to be appropriate for our needs.
- **Data Cleaning** - Dealing with Null and categorical data using:
    - Label Encoding for categorical variables
    - Converting Location column into State, Country and Area
    - Handling Null values and outliers
- **Feature Engineering** - Performed Feature Engineering to select relevant columns that have high correlation.
- **Text Analytics** - Most of the data is textual hence performed the following Text Analytics to deal with the textual content:
    - **Standardizing Text:** Handling and replacing all Special characters and punctuation marks
    - **Stemming:** This forms groups words having similar meanings.
    - **Lowercase Conversion:** This converts all uppercase letters to lowercase.
    - **Tokenization:** This converts text into tokens before converting it into vectors.
    - **Stop words Removal:** This removes all the "Stop words" such as able, to, from, has, me, myself etc.
- **Text Pre-processing** – Used **Count Vectorization** and **TD-IDF** to raw textual matter into matrix

Natural Language Processing (Write more and add images if u can)

Natural Language Processing is a way to process that textual data and turn it into numerical values or categorical values that you can use to model text.

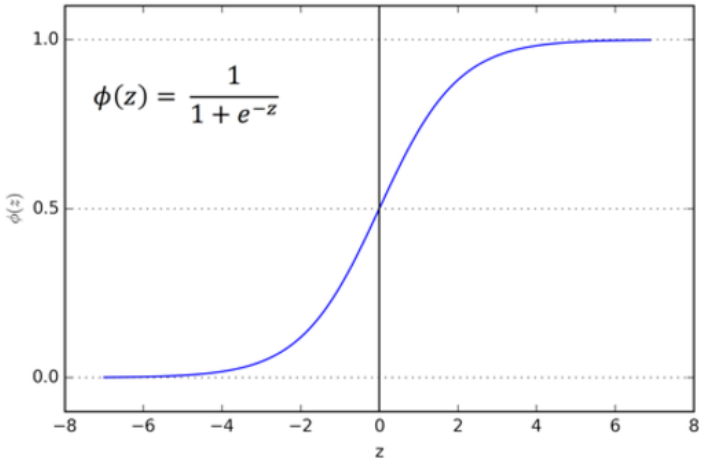
The aim of the project is to filter out all the fake jobs and perform exploratory data analysis to get more insights into such fake postings using Natural Language Processing.

- After treating the problem of Feature Engineering such as oversampling, Encoding the categorical values, handling null values, we move on to the converting text into numerical form to then perform the prediction part of the problem.
- We implement Count Vectorizer (bag of words) in order to calculate the occurrence of different words.
- On the basis of how frequently a word appears in the text, we calculate the Term Frequency and Inverse Document Frequency, which quantifies the importance of a string representation (words, phrases, lemmas) in our document.
- Then through Stemming and Lemmatization the dataset is transformed into the raw form and Machine Learning models are applied over the dataset.

Data Mining Models (left)

Logistics Regression

Logistic regression is a classification algorithm used for predictive analytics based on the concept of probability. Logistic regression uses complex cost functions. This function is called the "sigmoid function". The logistic regression hypothesis tends to limit the cost function between 0 and 1. Therefore, a linear function can have a value greater than 1 or a value less than 0 and cannot be represented.



KNN

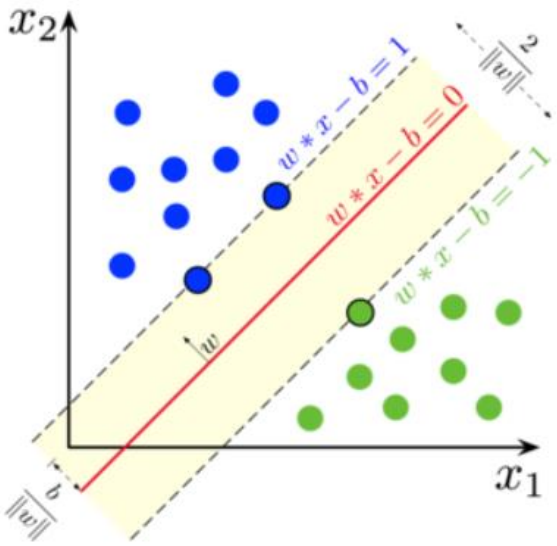
In the classification problem, the K-nearest neighbor algorithm essentially said that for a given value of K algorithm will find the K nearest neighbor of data point and then it will assign the class to these data point by having the class which has the highest number of data points out of all classes of K neighbors.

For distance metrics, we will use the Euclidean metric. Whose formula for distance is as follows:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \ldots + (x_n - x'_n)^2}$$

SVM

A Support Vector Machine or SVM is a machine learning algorithm that looks at data and sorts it into one of two categories and is usually used in solving binary classification problems.



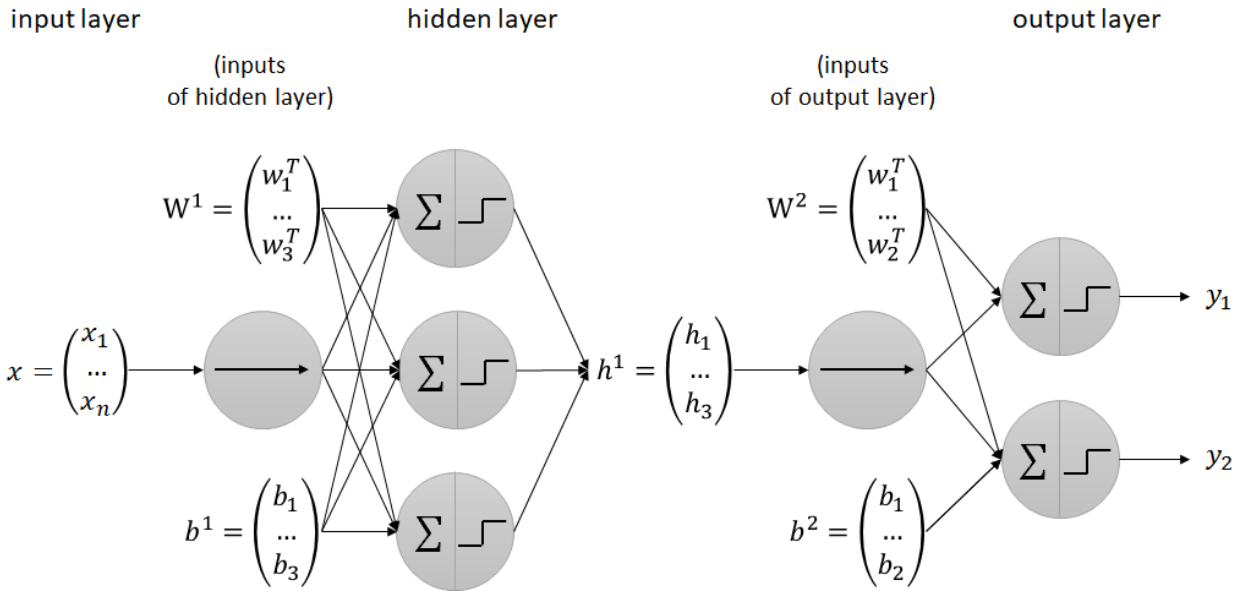
Random Forest

Random forests (RF) construct many individual decision trees at training. Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. Predictions from all trees are pooled to make the final prediction and we have used the mode for splitting as Gini Impurity.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a composition of an input layer, at least one hidden layer of LTUs and an output layer of LTUs. A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN).



Performance Evaluation

We performed modelling using Logistic Regression, KNN, SVC, MLP NN and Random Forest and evaluated various metrics like AUC ROC Score, F1 Score and Accuracy to help evaluate performances of the models after Hyper-Parameter Tuning.

Classification Reports

Logistic regression

	precision	recall	f1-score
0.0	0.86	0.90	0.88
1.0	0.78	0.70	0.74
accuracy			0.83
macro avg	0.82	0.80	0.81
weighted avg	0.83	0.83	0.83

KNN Gridsearch

	precision	recall	f1-score
0.0	1.00	0.82	0.90
1.0	0.73	1.00	0.84
accuracy			0.88
macro avg	0.87	0.91	0.87
weighted avg	0.91	0.88	0.88

Random Forest Classifier

	precision	recall	f1-score
0.0	1.00	0.99	1.00
1.0	0.99	1.00	0.99
accuracy			1.00
macro avg	0.99	1.00	1.00
weighted avg	1.00	1.00	1.00

Neural Network - Multi-Layer Perception

	precision	recall	f1-score
0.0	0.99	0.94	0.97
1.0	0.89	0.99	0.94
accuracy			0.96
macro avg	0.94	0.96	0.95
weighted avg	0.96	0.96	0.96

Confusion Matrix

Logistic regression

```
[1487,  164]
[ 248,  577]
```

KNN Gridsearch

```
[1348,  303]
[   0,  825]
```

Random Forest Classifier

```
[1640,   11]
[   0,  825]
```

Neural Network - Multi-Layer Perception

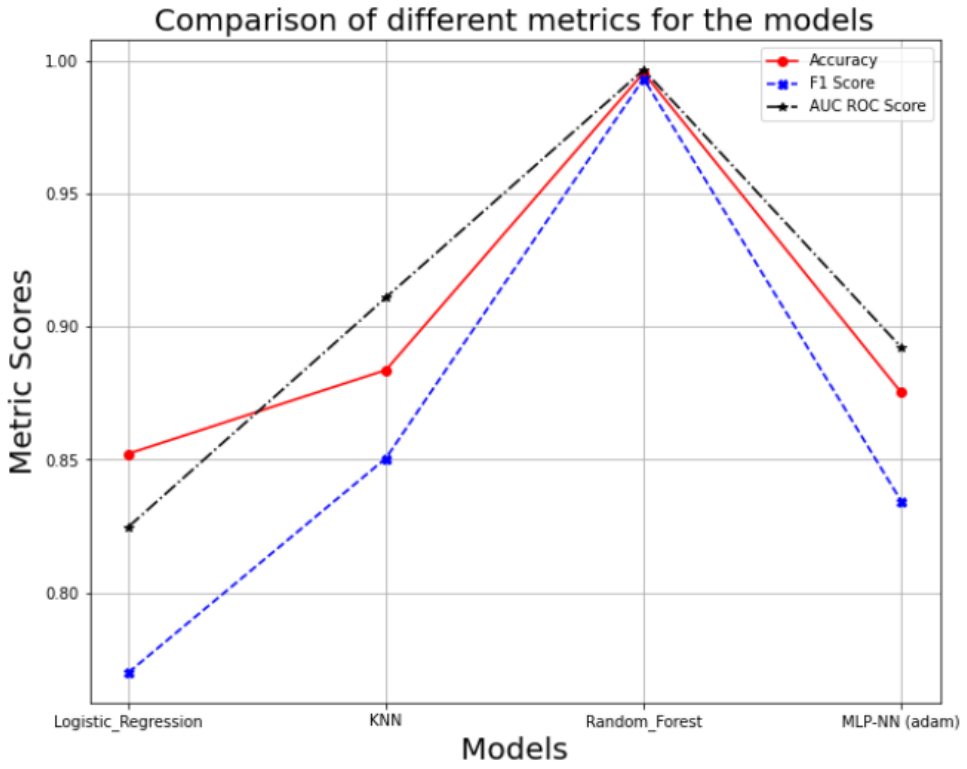
```
[1551,  100]
[   8,  817]
```

Comparision

We performed modelling using Logistic Regression, KNN, SVC, MLP NN and Random Forest and evaluated various metrics like AUC ROC Score, F1 Score and Accuracy to help evaluate performances of the models.

Model/ Metrices	AUC ROC Score	F1 Score	Accuracy Score	HyperParameters Tuned
Logistic Regression	0.8067	0.7369	0.8336	Penalty : l2 C : 0.1
KNN	0.9082	0.8449	0.8776	N_neighbours : 19
SVM	Not Cost Effective because it required a large computation power			Kernels : Linear & RBF
Random Forest	0.9967	0.9934	0.9956	N_estimators : 14
Multi-Layer Perception	0.9649	0.938	0.9564	hidden_layer_size: (100, 50, 30) max_iter : 1000

AUC – ROC Metric



- From this graph we can observe that values of Accuracy, F1 Score and AUC ROC Curve is Maximum for Random Forest as compared to rest of the models.
- The Random Forest outperforms all the other models in all three-evaluation metrics.
- Additionally, it doesn't require high computation power thus making it the best model for prognostication.

Project Results

We observe the best performing model was Random Forest on the testing dataset. This was due to the fact that the data was mostly textual and descriptive data hence since the model is suitable for dealing with the high dimensional noisy data, it works better than the rest.

Eval metrics for Random Forest

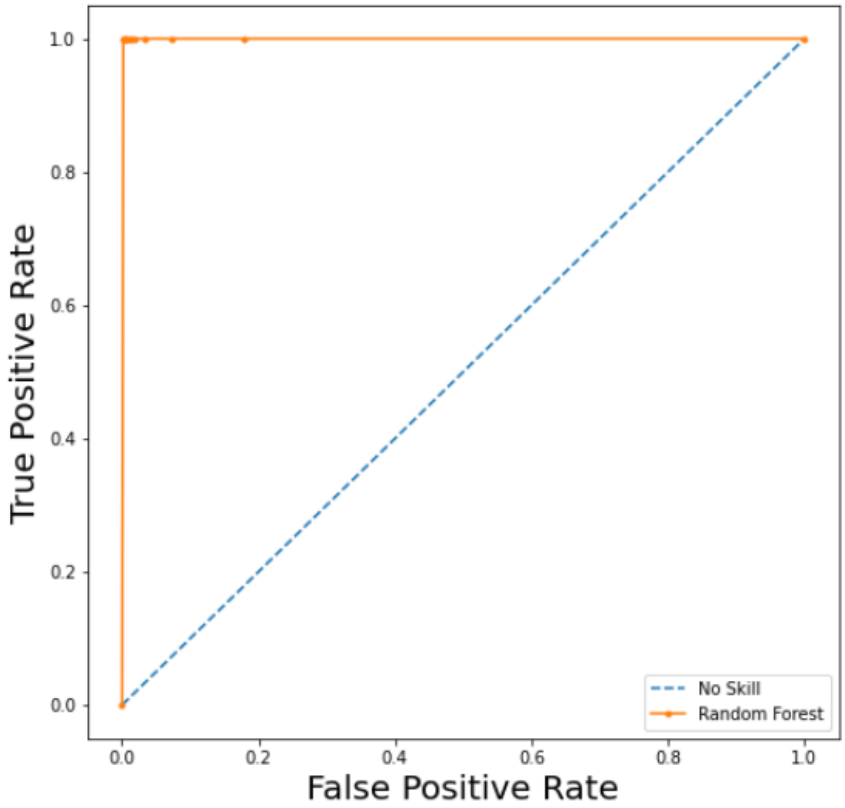
F1 Score: 0.9916

AUC ROC Score: 0.9966

Accuracy: 0.9956

	Real	Fraudulent
Real	1562	89
Fraudulent	122	703

Confusion Matrix



AUR ROC Score

Conclusion

- The Random Forest model performs better than the established benchmark of the baseline model as can be observed by the ROC AUC curve.
- The textual data is pre-processed to generate optimal results, and relevant numerical fields are chosen as well.
- The dataset that is used in this project is very unbalanced, so we performed oversampling to reduce the bias that the machine learning model has towards the dominant class.
- In the current economic scenario, there is a desperation to find a job as quickly as possible, but one should be vary of such scammers. This issue is a moral responsibility of the job-seeking community to help one another and report such fake postings if found to the concerned authorities.

Future Impact of Project Outcomes

- Live web scraping of job postings to speed up prediction time and minimize the impact of fraudulent scam.
- Creating a visual dashboard that would classify the jobs posted on various job sites such as LinkedIn, Indeed etc
- We can also create a visual dashboard for the display of key features or words that are common to most fraudulent jobs and how these features and their use by scammers changes over time.