**PROJECT PROPOSAL (Apogee)**

**Real or Fake job posting prediction**

**Team members:**

Manish Boyina – U59955856

Suryamouli Rajendra Varma Kathari – U69901705

Amit Gaur – U87764295

Monisha Lanka – U83830101

Prasanna Eddala – U24108900

**Problem Statement:**

Employment fraud is becoming more prevalent. According to CNBC, there were twice as many employment frauds in 2018 than there were in 2017. The state of the market today has resulted in substantial unemployment. The availability of work has been drastically reduced, and many people have lost their jobs as a result of economic hardship and the coronavirus's effects. Such a situation offers con artists the ideal opening. Due to a rare incidence, many people are becoming victims of scammers who feed on their despair. Most con artists use this technique to obtain personal information from their victims. Address, bank account information, social security number, and other personal information are examples. As a student, I have encountered several of these fraudulent emails. Users are offered a highly lucrative job opportunity by the con artists, who then demand payment. Or they demand money from the job seeker in exchange for the promise of employment. This is a dangerous problem that can be addressed through Machine Learning techniques.

The goal of this project is to develop a classifier that can distinguish between fake and real jobs. The final result will be evaluated based on different models. Both text and numeric features are included in the data, thus one model will be applied to the text data and the other to the numeric data. The result will be a synthesis of the two. The final model will include all relevant data from job postings and generate a decision on whether the job is real or not.

**Data source:**

The data collected and research done by the University of the Aegean. This research was done in 2020 after the covid pandemic. The data is collected from different sources and preprocessed by the researchers.

Reference link: [http://emscad.samos.aegean.gr/](http://emscad.samos.aegean.gr/%20)

Kaggle: <https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction>

**Data Information:**

The dataset consists of 17,880 observations and 18 features. The data is a combination of integer, binary and textual datatypes. The predictor variable is Fraudulent which is Boolean with 0 as real and 1 as fake. A brief definition of the variables is given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Datatype** | **Description** |
| 1 | job\_id | int | Identification number is given to each job posting |
| 2 | title | text | A name that describes the position or job |
| 3 | location | text | Information about where the job is located |
| 4 | department | text | Information about the department this job is offered by |
| 5 | salary\_range | text | Expected salary range |
| 6 | company\_profile | text | Information about the company |
| 7 | description | text | A brief description about the position offered |
| 8 | requirements | text | Pre-requisites to qualify for the job |
| 9 | benefits | text | Benefits provided by the job |
| 10 | telecommuting | boolean | Is work from home or remote work allowed |
| 11 | has\_company\_logo | boolean | Does the job posting have a company logo |
| 12 | has\_questions | boolean | Does the job posting have any questions |
| ­­13 | employment\_type | text | 5 categories – Full-time, part-time, contract, temporary and other |
| 14 | required\_experience | text | Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive, or Not Applicable |
| ­15 | required\_education | text | 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 | text | The industry the job posting is relevant to |
| 17 | Function | text | The umbrella term to determining a job’s functionality |
| 18 | Fraudulent | boolean | The target variable ◊ 0: Real, 1: Fake |

There are more the 0 in the predictor variable than 1. It clearly indicates that the dataset is an imbalanced dataset. We have to perform some sampling techniques to balance the data. Also, there are some null values in the data that need to be analyzed to remove or replace the values.

**Previous Evaluations:**

Previously, many competitors in Kaggle worked on this data and performed the analysis of the data, achieving good results using models like logistic regression, random forest, navies Bayes, and xgboost. But very few used the sampling technique on the data. Now let’s try to understand the data and perform our analysis.

**Model Evaluations:**

We will perform some model-building techniques for our data after preprocessing and data analysis. The following models are going to be implemented:

* Logistic regression
* Naive Bayes
* Decision tree classifier
* Random forest classifier
* XGBoost
* Adaboost

**Metrics:**

The models will be evaluated based on two metrics:

1. **Accuracy**: This metric is defined by this formula -

[equation](https://github.com/Anshupriya2694/Fake-Job-Posting-Prediction/blob/master/Images/accuracy.jpg)

As the formula suggests, this metric produces a ratio of all correctly categorized data points to all data points. This is particularly useful since we are trying to identify both real and fake jobs, unlike a scenario where only one category is important. There is however one drawback to this metric. Machine learning algorithms tend to favor dominant classes. Since our classes are highly unbalanced a high accuracy would only be a representative of how well our model is categorizing the negative class (real jobs).

1. **F1-Score:** The F1-score is a measure of a model’s accuracy on a dataset. The formula for this metric is –

[equation](https://github.com/Anshupriya2694/Fake-Job-Posting-Prediction/blob/master/Images/f1-score.jpg)

F1-score is used because in this scenario both false negatives and false positives are crucial. This model needs to identify both categories with the highest possible score since both have high costs associated with them.

**Applications:**

After we do model building for the data, we choose the best fit model and we can implement it as a product. Even though we can also present this model to the top job posting websites or companies they can deploy it into their servers to detect jobs fraudulently and those postings to prevent the applicants from scams.