BAN 620 PROJECT REPORT

TOPIC:

FAKE JOB POSTINGS

SHRIYA ARORA – sz9461

PALLAVI NAIR- fb4097

PRIYADARSHINI MADHUSUDANAN- hy1162

ADITYA MANE- yy5910

SAI KUMAR- lp2345

HARSHITHA ANANTHULA- pi3407

CAL STATE EAST BAY

Table of Contents

INTRODUCTION:

🔍 Analyzing Fake Job Postings: Combating Fraud through Data Mining Techniques

In the search for employment, job seekers often fall victim to fraud and scams. False job postings not only lead to financial losses but also pose risks such as identity theft and personal harm. The prevalence of employment fraud has been on the rise, with CNBC reporting a doubling of cases in 2018 compared to the previous year.

Amidst the current job market challenges, exacerbated by the COVID-19 pandemic, the desperation of job seekers has created an ideal opportunity for con artists. Exploiting this vulnerability, scammers deceive individuals by offering seemingly lucrative job opportunities, only to demand payment or personal information as part of their scheme.

To address this issue, our project focused on analyzing a comprehensive dataset sourced from Kaggle. With approximately 18,000 records and 18 variables, the dataset includes diverse job postings from different countries and positions. By leveraging data mining techniques, including preprocessing, exploratory analysis, machine learning, and natural language processing (NLP), we aimed to identify patterns and develop predictive models.

By understanding the characteristics of fake job postings and uncovering hidden insights, we strive to contribute to the prevention and detection of fraudulent activities. Our goal is to protect job seekers from falling victim to scams, ensuring a safer and more secure employment landscape.

GOAL:

To address the growing issue of fraudulent job ads on internet platforms, we are developing a fake job detection model, these models are made to instantly recognize and flag questionable or fraudulent job advertisements and prevent job seekers from falling prey to such scams by identifying and filtering out fake postings.

DATA ANALYSIS:

Initially, the data has 18 attributes in which the target variable is considered as the fraudulent variable.

The data has 17,880 records in total.

```
In [3]: job_df.shape
Out[3]: (17880, 18)
```

The names of all the variables from the dataset are listed below:

```
In [30]: job_df.columns

Out[30]: Index(['job_id', 'title', 'location', 'department', 'salary_range',
                'company_profile', 'description', 'requirements', 'benefits',
                'telecommuting', 'has_company_logo', 'has_questions', 'employment_type',
                'required_experience', 'required_education', 'industry', 'function',
                'fraudulent'],
               dtype='object')
```

Step1 - Initial Review:

- After the initial assessment, it could be observed (Figure 1 & 2) that variables such as Salary range (84%) and Department (65%) showed the highest number of missing values.

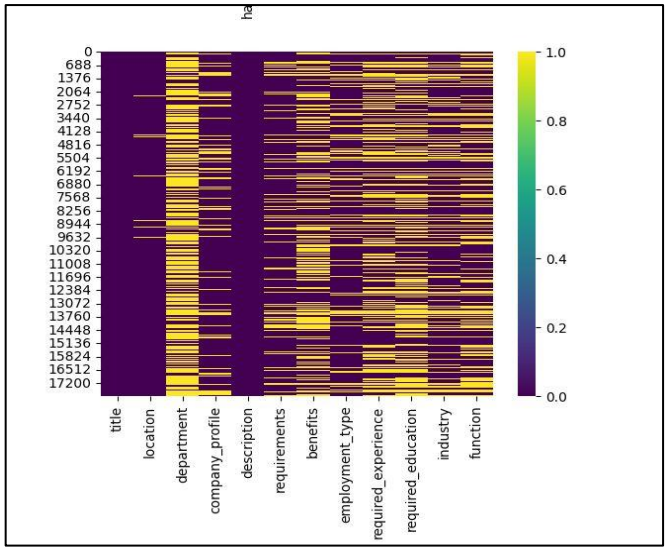

Figure 1: Heat Map showing the missing values.

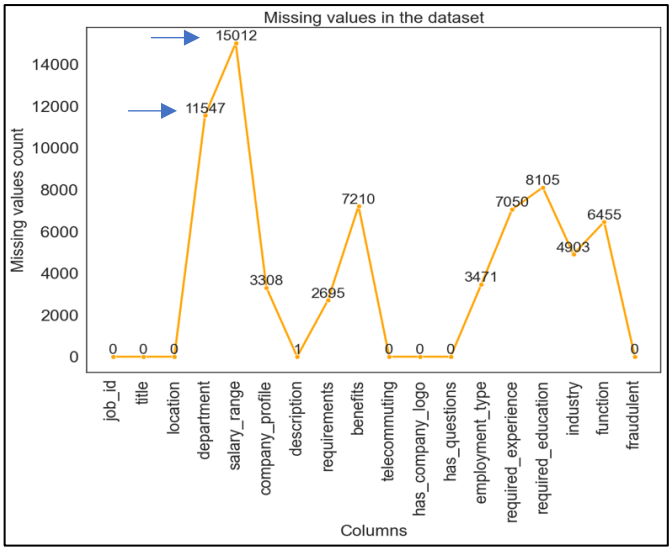

Figure 2: Missing values count

- As most of the data is categorical, we began exploring data from the location perspective. We can notice that the United States is a clear winner with 60% of the job postings in this dataset.

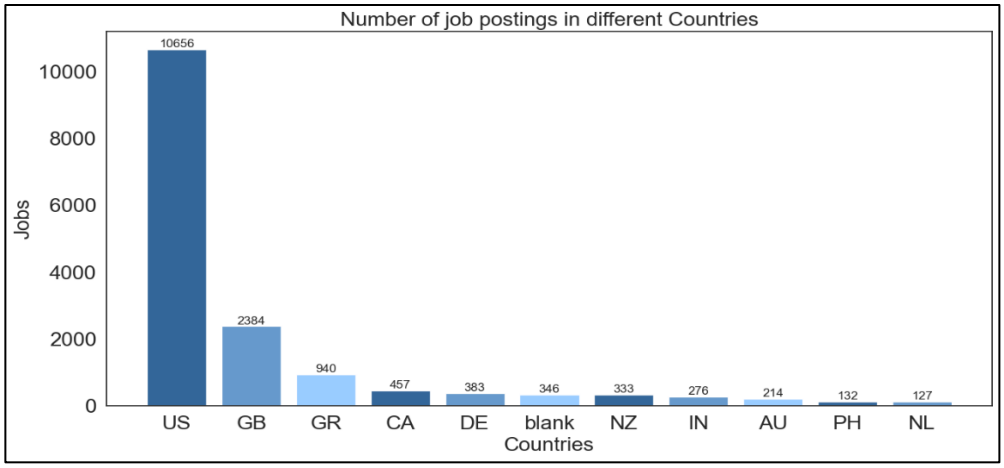

Figure 3: Number of Job postings by countries

- Furthermore, we can see the count of Fraudulent data (Figure 4), which is 5% of the jobs in the dataset versus 95% of the jobs in the dataset are real, which shows that the dataset is highly imbalanced.

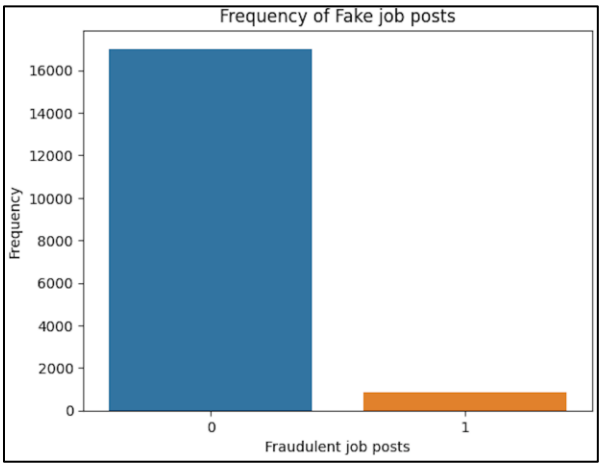

Figure 4: Frequency of Fake Jobs

Step 2- Data Preprocessing:

- Considering the huge number of missing values and as the dataset has textual data in the majority, the Pre-processing step is very crucial for building and implementing our model We have focused on three major details here:

```
[8]: job_df.isnull().sum()

[8]: title                    0
     location               346
     department           11546
     company_profile       3307
     description              0
     requirements          2694
     benefits              7209
     telecommuting            0
     has_company_logo         0
     has_questions            0
     employment_type       3470
     required_experience   7049
     required_education    8104
     industry              4902
     function              6454
     fraudulent               0
     dtype: int64
```

```
In [16]: df2.isnull().sum()

Out[16]: title                    0
         location                19
         department             531
         salary_range           643
         company_profile        587
         description              1
         requirements           154
         benefits               364
         telecommuting            0
         has_company_logo         0
         has_questions            0
         employment_type        241
         required_experience    435
         required_education     451
         industry               275
         function               337
         fraudulent               0
         dtype: int64
```

- The above tables give information about the non-null values and the data types of each variable in the entire data set vs the missing values in the fraudulent jobs.

Treating missing values:

- In the first step of data cleaning, we dropped the *job_id* and the *salary range* columns as job_id is the sequential order of numbers and the salary_range column has a greater number of null values.
- We treated all the missing values using “missing” text so that we improve data quality, minimize bias, and enable complete case analysis, all of them give us accurate imputation. These benefits contribute to more reliable and insightful data analysis and modeling outcomes.

Encoding the categorical variables:

- Preprocessing the categorical variables is a necessary step. We need to convert the categorical variables to numbers such that the model can understand and extract valuable information. The columns that were encoded are *Employment type, Required experience, Required education, Industry, Function, Country, and Department.*

Text Mining techniques:

- Our data majorly focuses on text variables; hence text mining did play an important part in model development, the techniques we used are:
- Changing all the characters to lowercase.
- Removal of special characters.

- Tokenization - A process of breaking up a given text into units called tokens. Tokens can be individual words, phrases, or even whole sentences. For example, consider the sentence: “Never give up”. The most common way of forming tokens is based on space. Assuming space as a delimiter, the tokenization of the sentence results in 3 tokens – Never-give-up. As each token is a word, it becomes an example of Word tokenization.
- Stop word removal: One of the most used preprocessing steps across different NLP applications. The idea is to remove the words that occur commonly across all the documents.
- Lemmatization: It considers the context and converts the word to its meaningful base form, which is called Lemma. For instance, stemming the word 'Caring' would return to 'Car'.
- Vectorization: Converting words to vectors, or word vectorization, is a natural language processing (NLP) process. The process uses language models to map words into vector space. A vector space represents each word by a vector of real numbers. It also allows words with similar meanings to have similar representations.
- We have preprocessed and vectorized the following text variables: 'title', 'company_profile', 'description', 'requirements', 'benefits'.

For our project, we have considered title as a text variable and not a categorical variable because 11,233 records in the dataset had unique values for the variable ‘title’. (63% of the entire data set had unique titles)

Step 3- Exploratory data analysis (EDA): Uncovering Insights and Enhancing Data Mining

- In our project, Exploratory Data Analysis (EDA) played a crucial role in understanding the dataset and setting the stage for subsequent data mining tasks. By delving into the data, we gained a comprehensive understanding of its characteristics, identified potential challenges, and paved the way for further analysis.
- EDA allowed us to examine the distribution of variables, detect outliers, explore relationships between features, and uncover hidden patterns or trends. By visualizing the data through charts, graphs, and statistical measures, we gained valuable insights into the dataset's structure and content.
- Furthermore, EDA helped us identify and address data quality issues, such as missing values, inconsistent formatting, or anomalies. By cleansing and preprocessing the data, we ensured its reliability and prepared it for subsequent modeling and analysis.
- By leveraging EDA techniques, we were able to extract meaningful insights, make informed decisions, and develop robust data mining models. This process not only enhanced our understanding of the dataset but also laid the foundation for uncovering valuable patterns and predicting outcomes.

- The figures (5,6,7) below portray an overall picture of the total number of Jobs postings in each of the below-mentioned features such as ***employment, experience, and education***. Moreover, it also gives us important information of fraudulent jobs pertaining to these specific features. We can interpret that the main targets in terms of fraudulent jobs were seen in the full-time category, entry-level experience, and high school education.

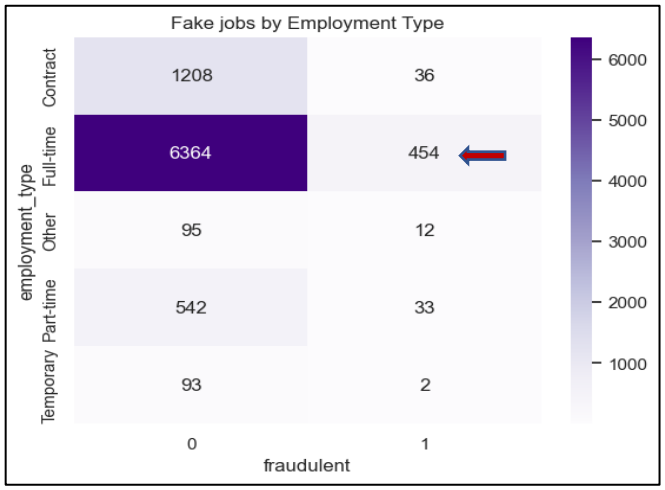

Figure 5: Fraudulent Jobs by Employment

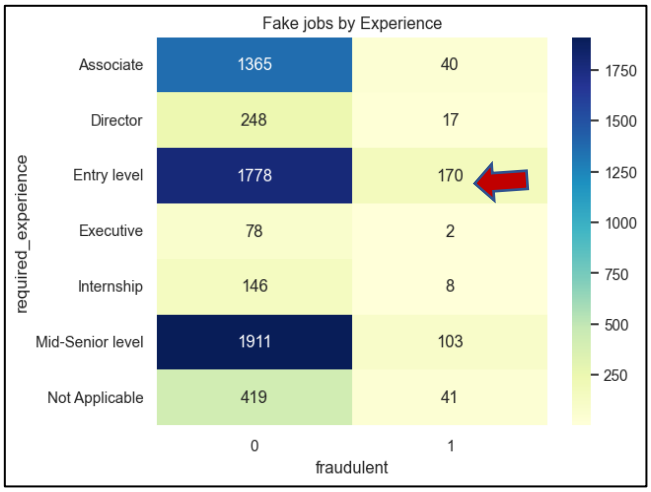

Figure 6: Fraudulent Jobs by Experience

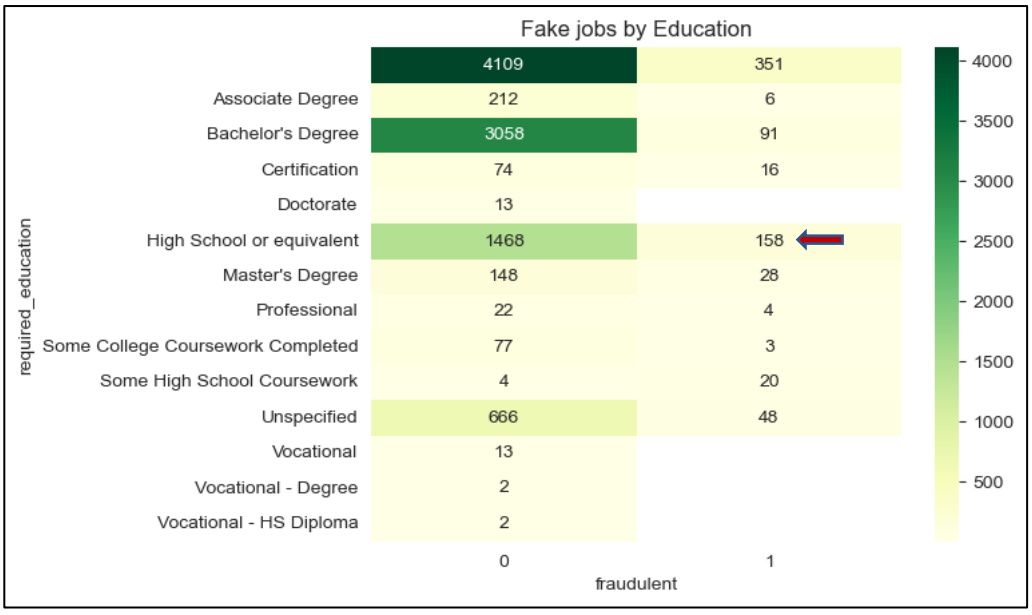

Figure 5: Fraudulent Jobs by Education

- As previously noted, we could see that the US region (60%) was dominating the number of job postings in our dataset, hence for the purpose of visualization, we have concentrated on the US region and further drilled down into the State and City wise breakdowns.

- In addition, according to Figure 8 below, ***California*** tops the list of states with the highest number of job postings followed by ***New York and Texas.***

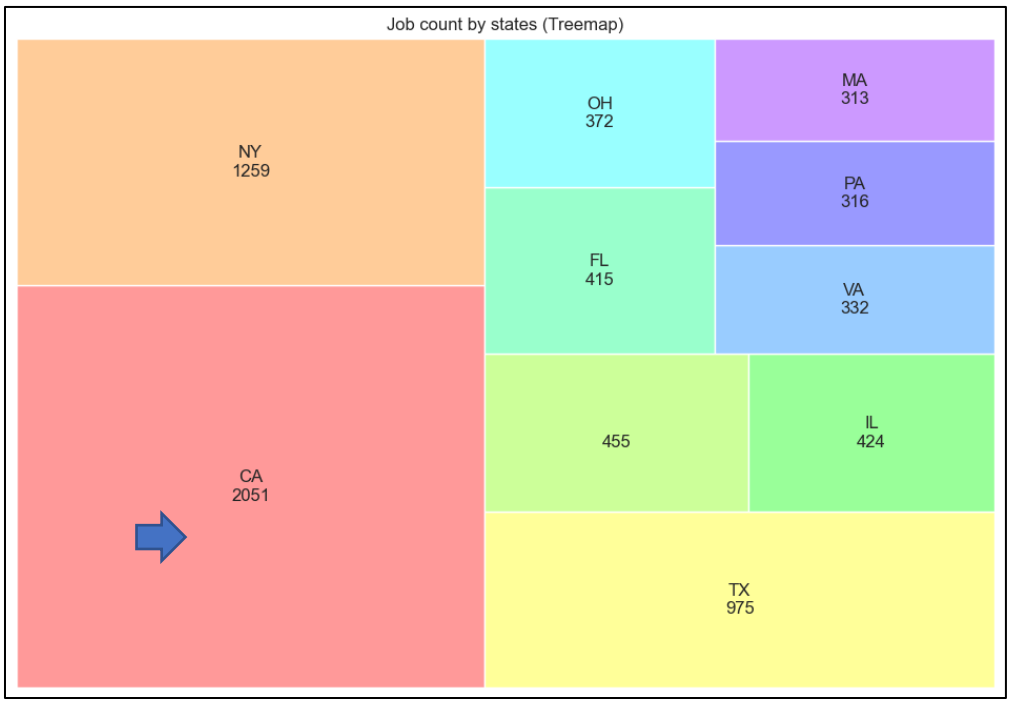

Figure 8: Job postings by States

- To explore further, the next bar plot (Figure 9) shows that the highest possibility of fake jobs is in Texas followed by California.
- To further drill down at a city level, we have calculated the fake to real jobs including the state and cities. The Fake to Real jobs ratio (Figure 10) that stood out was Bakersfield, California ((15:1) followed by Dallas, Texas (12:1), hence showing that any job postings from these areas in specific might have a higher chance of being fraudulent.

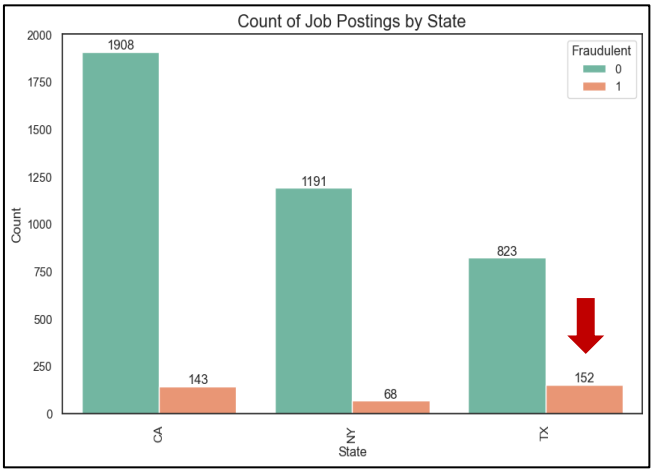

Figure 9: Top 3 States with Fake Jobs

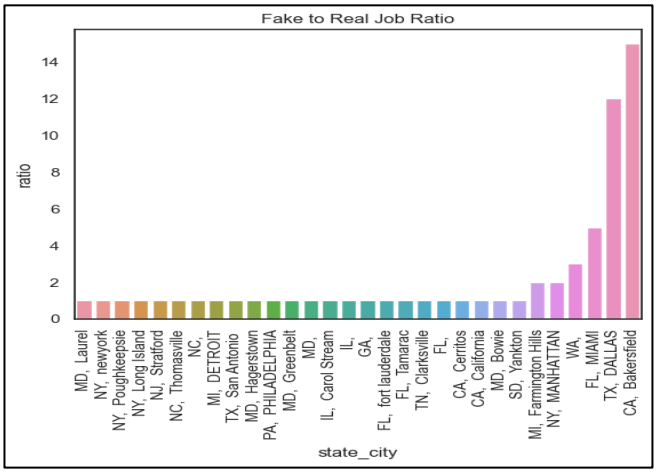

Figure 10: Fake to Real Job Ratio

Step 4: Exploring different data mining models:

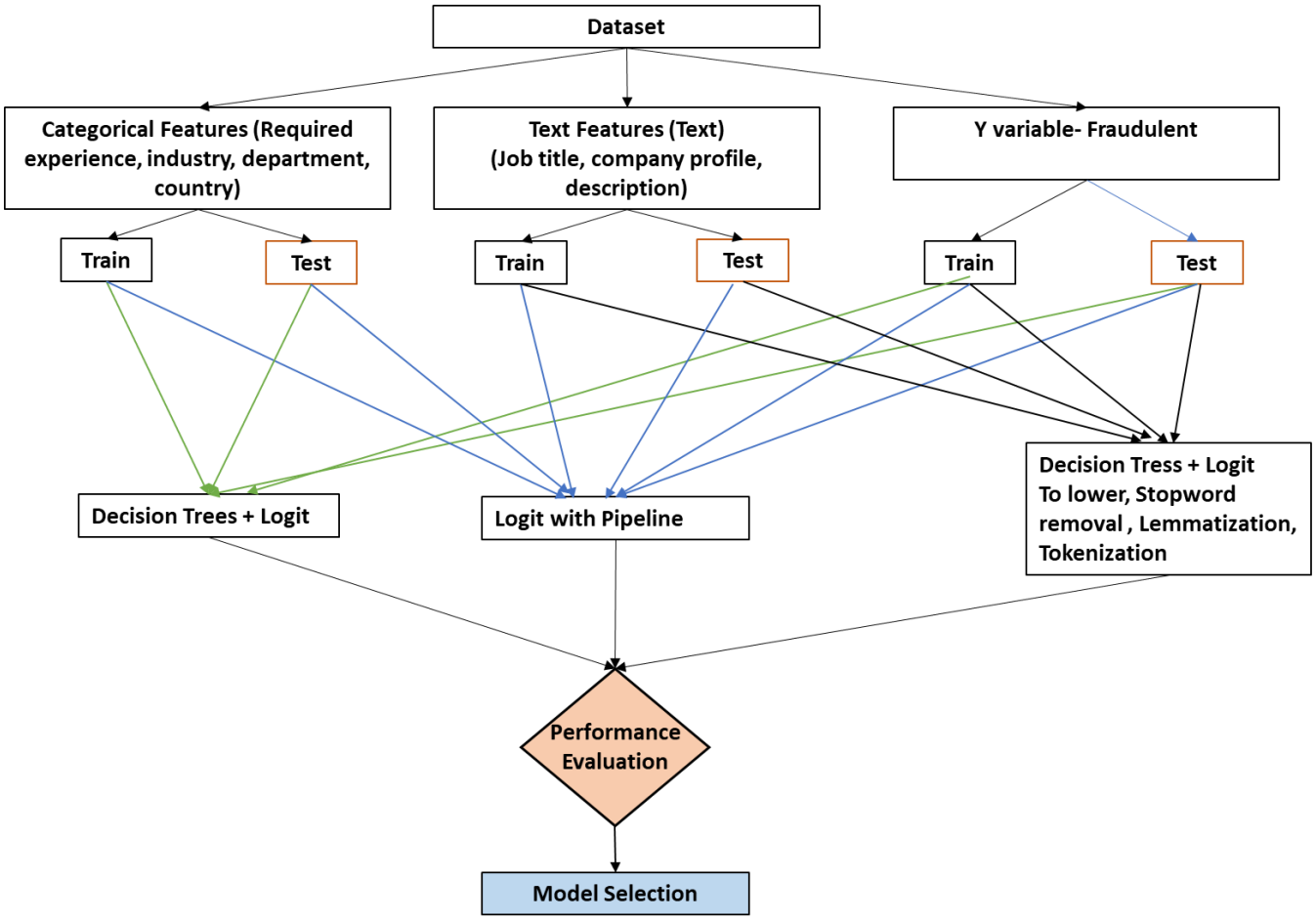

Figure 11. Data models

1. Considering categorical variables:

We have considered 'telecommuting', 'has_company_logo', 'has_questions', 'employment_type', 'required_experience','required_education', 'industry', 'function' ,'department', 'Country' as the predictor variables and fraudulent as the outcome variable and built the following models.

a. Decision tree classifier:

Using only the categorical variables from the decision tree we achieved an accuracy of **95.81%** and from the confusion matrix, we observed sensitivity to be low. The actual fake jobs are more however, the model can predict a low number of fake jobs.

```
Confusion Matrix (Accuracy 0.9581)

       Prediction
Actual    0    1
     0 5088   16
     1  209   51
```

Figure 12: Confusion Matrix of Decision tree

From the decision tree, the best class tree is chosen, and its accuracy and the confusion matrix are shown below:

```
Confusion Matrix (Accuracy 0.9616)

       Prediction
Actual    0    1
     0 5078   26
     1  180   80
```

Figure 13: Confusion Matrix of Best Class tree

b. Random Forest Classifier:

Using the random forest classifier, we have achieved an accuracy of **97.20%** and the confusion matrix is shown below:

```
Confusion Matrix (Accuracy 0.9720)

       Prediction
Actual    0    1
     0 5078   26
     1  124  136
```

Figure 14: Confusion Matrix of Random Forest

c. Gradient Boosting Classifier:

```
Confusion Matrix (Accuracy 0.9623)

       Prediction
Actual    0    1
     0 5089   15
     1  187   73
```

Figure 15: Confusion Matrix of Booster Trees

d. Logistic regression model:

With logit model we achieved an accuracy of 95.73% but the confusion matrix states that the model predicted only 20% of the Fake jobs in the validation data

```
Confusion Matrix (Accuracy 0.9573)

       Prediction
Actual    0    1
     0 5081   23
     1  206   54
```

Figure 16: Confusion Matrix of Logistic Regression Model

2. Considering text variables:

We have considered 'title', 'company_profile', 'description', 'requirements', 'benefits' as the predictor variables and fraudulent as the outcome variable and built the following models.

As mentioned earlier, we preprocessed the text variables and deployed them in the models.

a. Logit model:

The logit model with all the text we have achieved an accuracy of 98.21% and the confusion matrix states that the model predicted only 65% of the Fake jobs in the validation data. Sensitivity is improved in this model.

```
Confusion Matrix (Accuracy 0.9821)

       Prediction
Actual    0    1
     0 5100    4
     1   92  168
```

Figure 17: Confusion Matrix of Logistic Regression Model with text variables

b. Decision Tree Classifier:

This model with the text data gave us an accuracy of 98.15 and a good sensitivity (67%) as well.

```
Confusion Matrix (Accuracy 0.9815)

       Prediction
Actual    0    1
     0 5091   13
     1   86  174
```

Figure 18: Confusion Matrix of Decision Tree Classifier with text variables

3. Considering both categorical and text variables:

We considered the important categorical variables ('has_questions', 'required_experience', 'industry', 'department', 'Country') and text variables ('title', 'company_profile', 'description', 'requirements', 'benefits') as predictors and ‘fraudulent’ as the outcome variable and built a Logistic Regression model.

a. Logit model:

During the exploration process and to gain more accuracy and best sensitivity, a further step considering both categorical and text variables in the logit model gained an accuracy of 98.40% with best sensitivity compared to all the above models.

Firstly, we preprocessed the text variables and vectorized them. Then, we considered only the most important categorical variables, encoded them using OneHotEncoder and added them to a pipeline. Then, we concatenated the vectorized text in the pipeline and provided it as input to the Logistic Regression model.

```
Confusion Matrix (Accuracy 0.9840)

       Prediction
Actual    0    1
     0 5082   22
     1   64  196
```

Figure 18: Confusion Matrix of Logistic Regression Model with text variables & Categorical variables

As observed in the confusion matrix, we obtained **an accuracy of 98.4% with 75% sensitivity in predicting Fake jobs.**

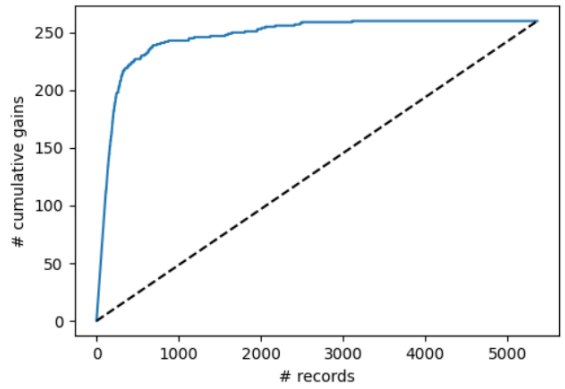

Figure 19: Cumulative Gains Chart

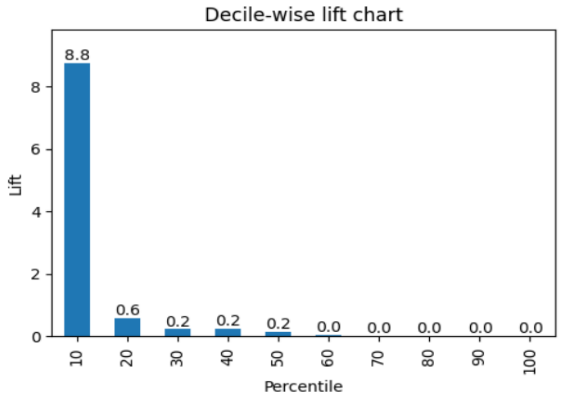

Figure 20: Decile-wise lift Chart

Plans moving forward:

Our team is committed to enhancing the accuracy of our fake job prediction model by leveraging advanced techniques. In our ongoing research, we plan to explore the following methods:

- Oversampling: By implementing oversampling techniques, we aim to address the class imbalance issue in our dataset. This approach involves generating synthetic samples of the minority class (i.e., fake jobs) to create a more balanced training set. This can help improve the model's ability to detect fake jobs accurately.

- BERT (Bidirectional Encoder Representations from Transformers): BERT is a powerful natural language processing model that captures contextual relationships in text data. By incorporating BERT into our model, we expect to enhance the understanding of text variables and improve the overall performance of our fake job prediction.

- LDA (Latent Dirichlet Allocation): LDA is a topic modeling technique that uncovers hidden topics within a collection of documents. By applying LDA to our text variables, we aim to extract meaningful topics related to job postings.

Through these explorations, we strive to gain a deeper understanding of the text variables and employ advanced methodologies to enhance the accuracy of our fake job prediction model.

Recommendations: *Identifying Real vs. Fake Job Postings-Key Factors to Consider*

When evaluating job postings for authenticity, it is crucial for job seekers to exercise caution and verify the credibility of the opportunity. Here are key factors to consider:

- Company Profile: Thoroughly examine the company profile provided in the job posting. If it is vague, incomplete, or missing entirely, it could be a red flag. Authentic job postings typically include detailed company information, such as a description, website, and contact details.
- Salary Range: Compare the salary range offered with the prevailing market trends for similar positions. If the salary appears exceptionally high or unrealistic, it may indicate a potential scam. Exercise caution and research average salary ranges for similar roles in your industry and location.
- Job Description and Title: Analyze the job description and title provided in the posting. Authentic job postings usually include comprehensive details about the role, responsibilities, and required qualifications. Cross-check this information with the open positions listed on the company's official website or other reputable job portals.

It's important to note that these factors are not exhaustive, and job seekers should exercise additional due diligence when assessing job postings. Here are some general tips:

- Verify the legitimacy of the company by researching its online presence, reviews, and reputation.
- Be cautious of job postings that demand upfront payment or require personal financial information.
- Look for consistent and professional communication throughout the hiring process.
- Trust your instincts and be wary of job opportunities that seem too good to be true.
- Remember, it's always recommended to cross-check multiple times, gather as much information as possible, and consult reliable sources to ensure the authenticity of job postings.