**Project Proposal – Group 6**

**Real/Fake Job Posting Prediction**

**Problem:**

The problem we aim to tackle in this project is the identification of fraudulent job postings. With the rise of online job platforms, the potential for harmful fake job listings has increased significantly, resulting in individuals falling victim to scams that may lead to the loss of personal information or financial resources. The goal of this project is to develop a machine learning model that can accurately classify job postings as either real or fake. By building such a model, we can provide insights into the characteristics that make a job posting fraudulent, offering a valuable tool for both job platforms and job seekers to ensure a safer job application process.

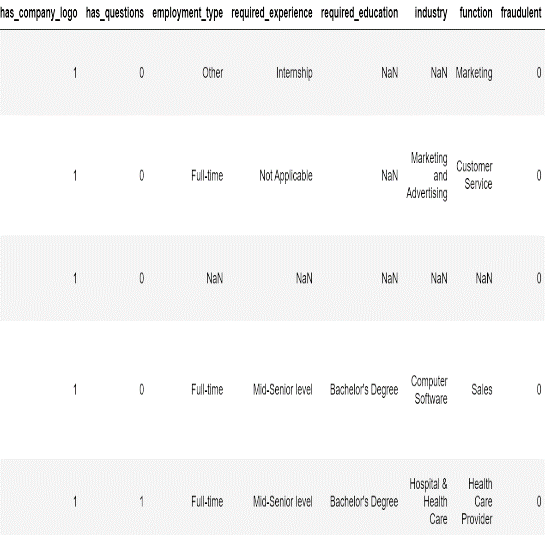
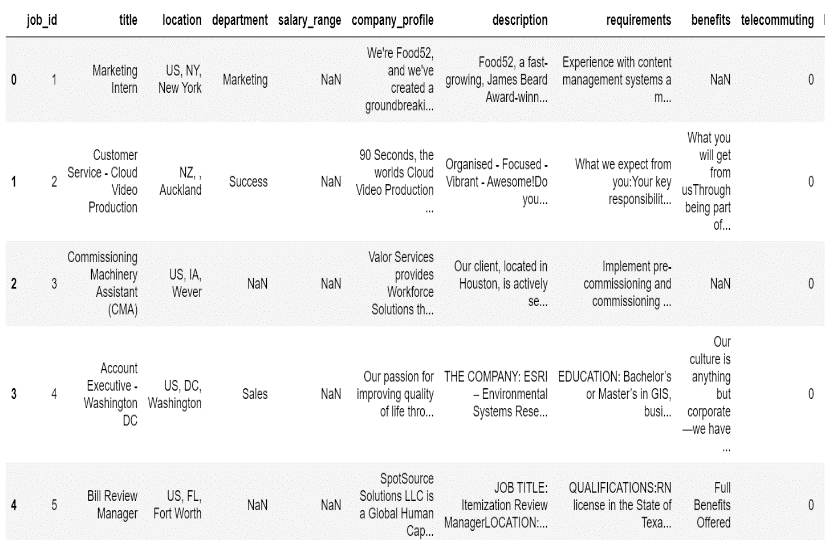
**Solution:**

The solution has several key benefits. For job platforms, integrating such a classification model can help filter out fraudulent listings before they are published, reducing the risk to users. For job seekers, this model provides a layer of protection, enabling them to navigate job listings with more confidence. From a business perspective, enhancing trust in online platforms can increase user engagement and trust in the service, making this solution not only helpful for individuals but also valuable for businesses and regulators concerned with job market integrity.

**Data:**

The dataset we will be using contains around 18,000 job descriptions, of which approximately 800 are labeled as fake. It includes both textual job descriptions and meta-features such as job title, company, location, and employment type. This rich combination of features allows us to explore patterns in both the descriptions and the accompanying details, helping us identify common traits associated with fake job postings. The dataset is accessed from Kaggle, providing a valuable resource for analysis and machine learning model training.

**Dataset-**[**https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data**](https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data)

Dataset Snapshot

**Techniques:**

For this project, we will perform extensive exploratory data analysis (EDA) to summarize the dataset and identify trends. We will preprocess the job descriptions using natural language processing (NLP) techniques, including tokenization, stop-word removal, and lemmatization, to prepare the text data for model training. Feature extraction will involve using both text features and meta-features, and we plan to apply machine learning models such as Random Forest, Logistic Regression, and Neural Networks for classification.

**Challenges:**

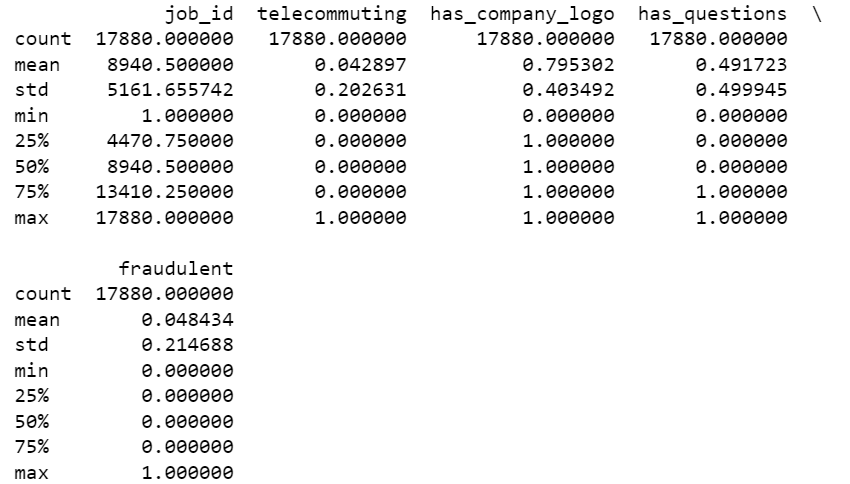
There are some anticipated challenges in this project. One of the primary difficulties will be the class imbalance, as only 800 of the 18,000 job postings are fake. This imbalance can make it difficult for the model to accurately predict fraudulent listings without overfitting. Additionally, extracting meaningful features from the text data will require careful preprocessing and experimentation with different techniques to ensure the model captures relevant information. Ensuring the model generalizes well to new data is also crucial to the success of this project, as it must be able to handle previously unseen job postings accurately.

**Research questions:**

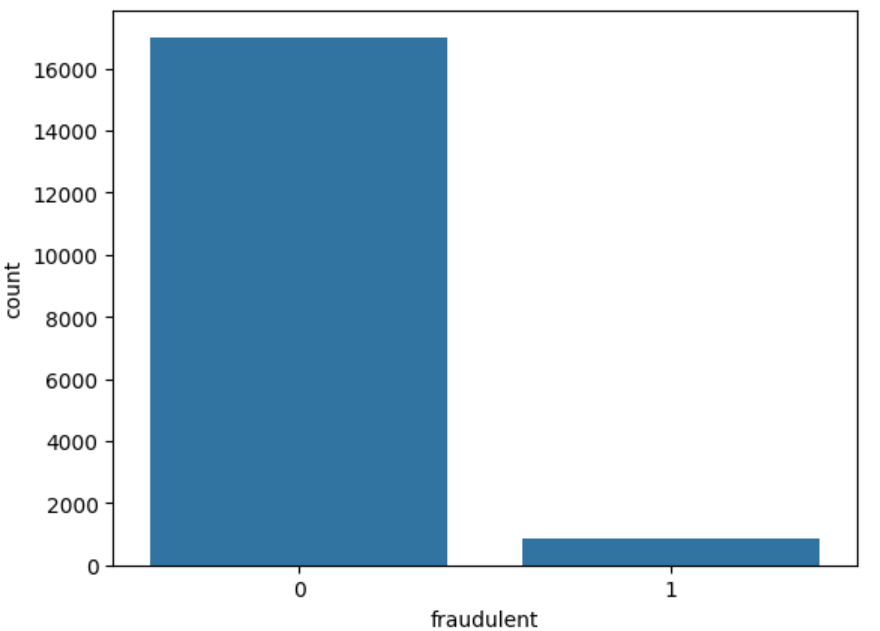
Throughout the project, we aim to investigate several key research questions. These include identifying the main features of both real and fake job postings, building a model that can predict fraudulent postings with high accuracy, and understanding the specific words and phrases most associated with fake job descriptions.

**Exploratory analysis:**

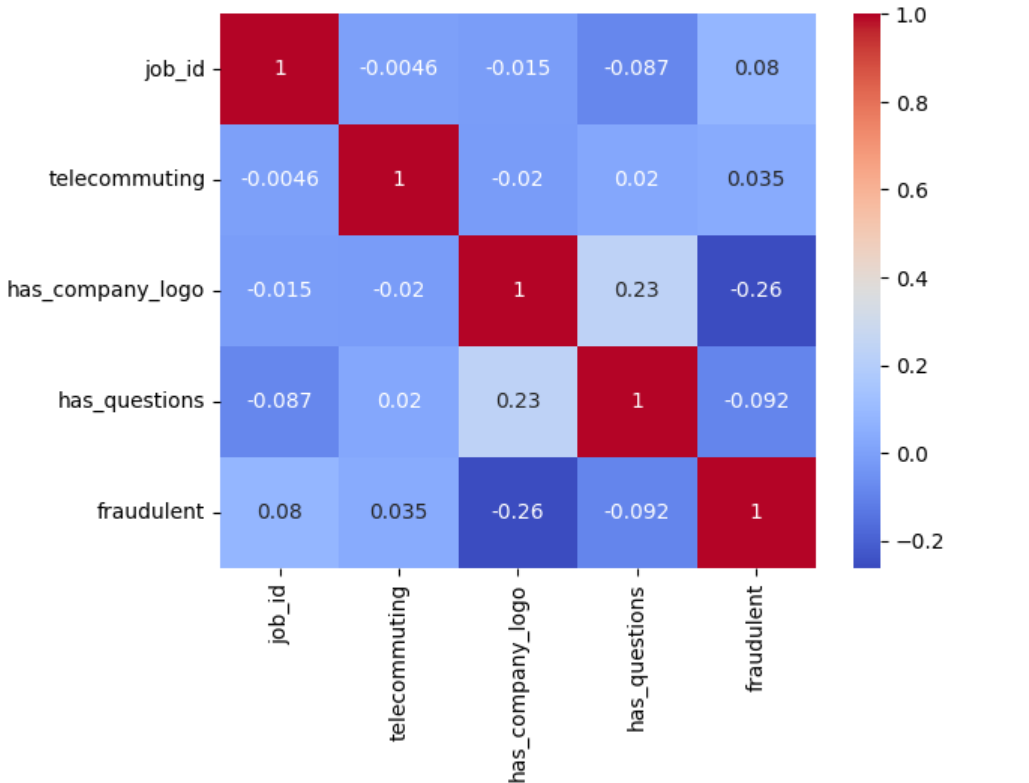
Here are some basic exploratory analysis, including statistical summaries and data visualizations, to better understand the dataset:



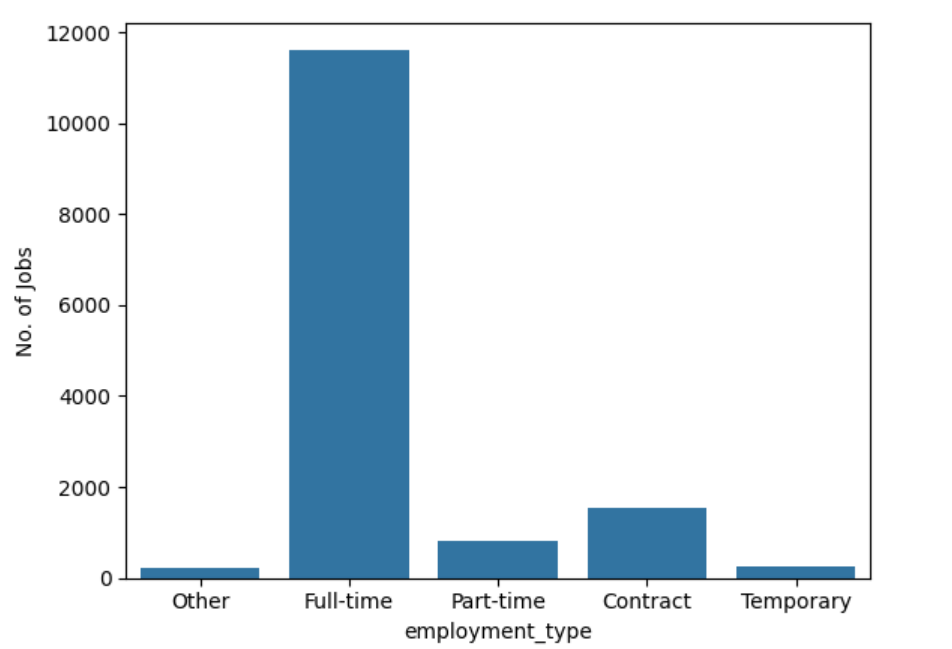
Statistical Summary of Job Posting Features



Count of Real Jobs (Labeled 0) vs Fake Jobs (Labeled 1) in the Dataset



Heatmap of Numerical Features



Count of Job Postings by Employment Type