Analyzing Real and Fake Job postings

Business Understanding

- **Understanding the Problem**

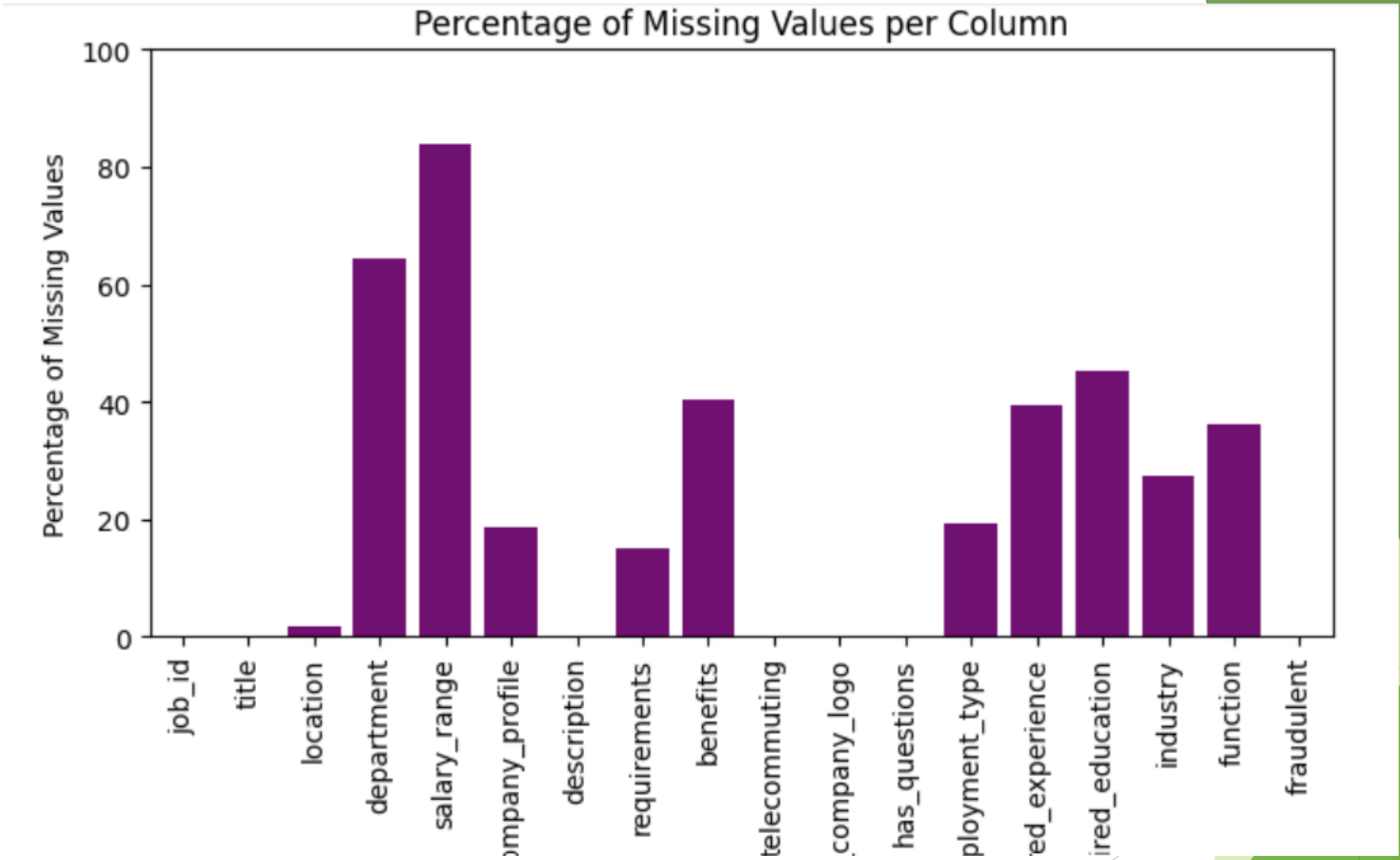
Fraudulent Job Postings: These present significant risks, including financial losses and identity theft.

- **Dataset Insights**: Contains 18,000 job descriptions with about 800 fraudulent, including text data and meta information.

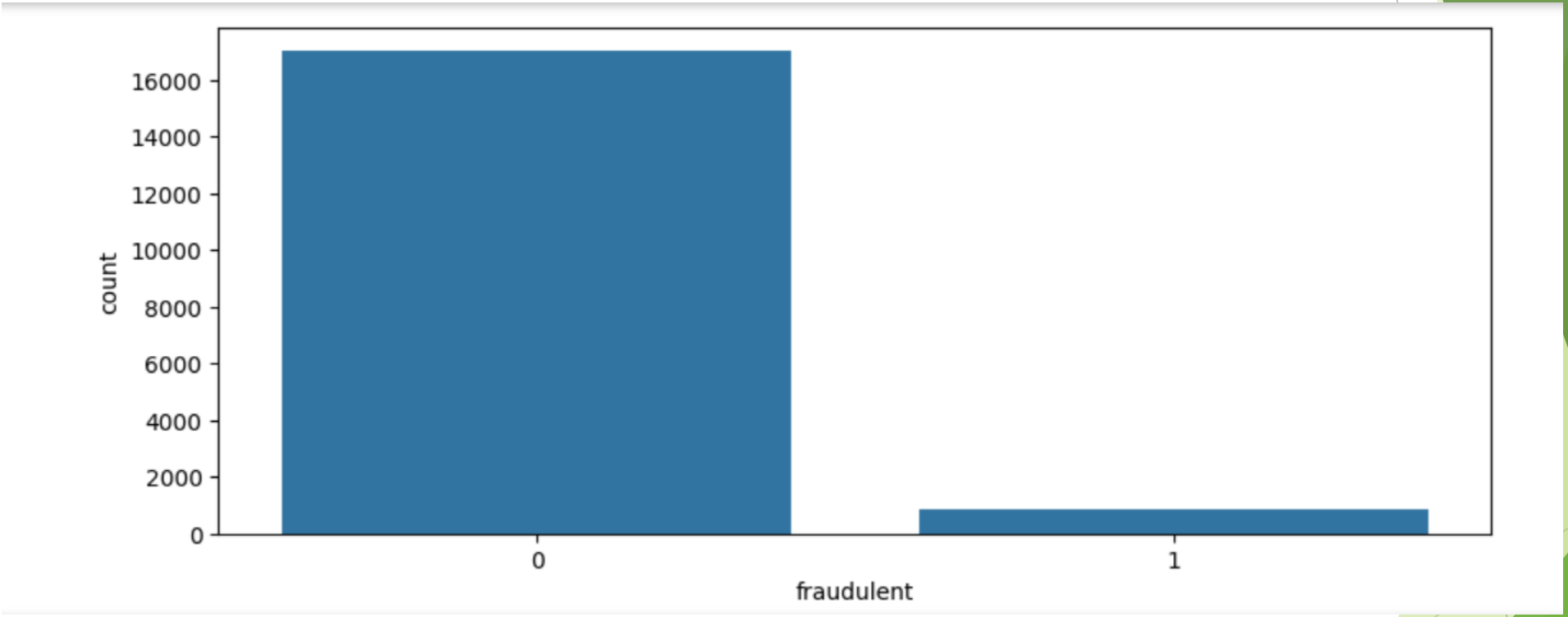
- **Project Objectives**
- **Predictive Modeling**: Aim to develop a classification model to distinguish between real and fake job postings.
- **Feature Identification**: Discover key traits indicative of fraudulent postings.

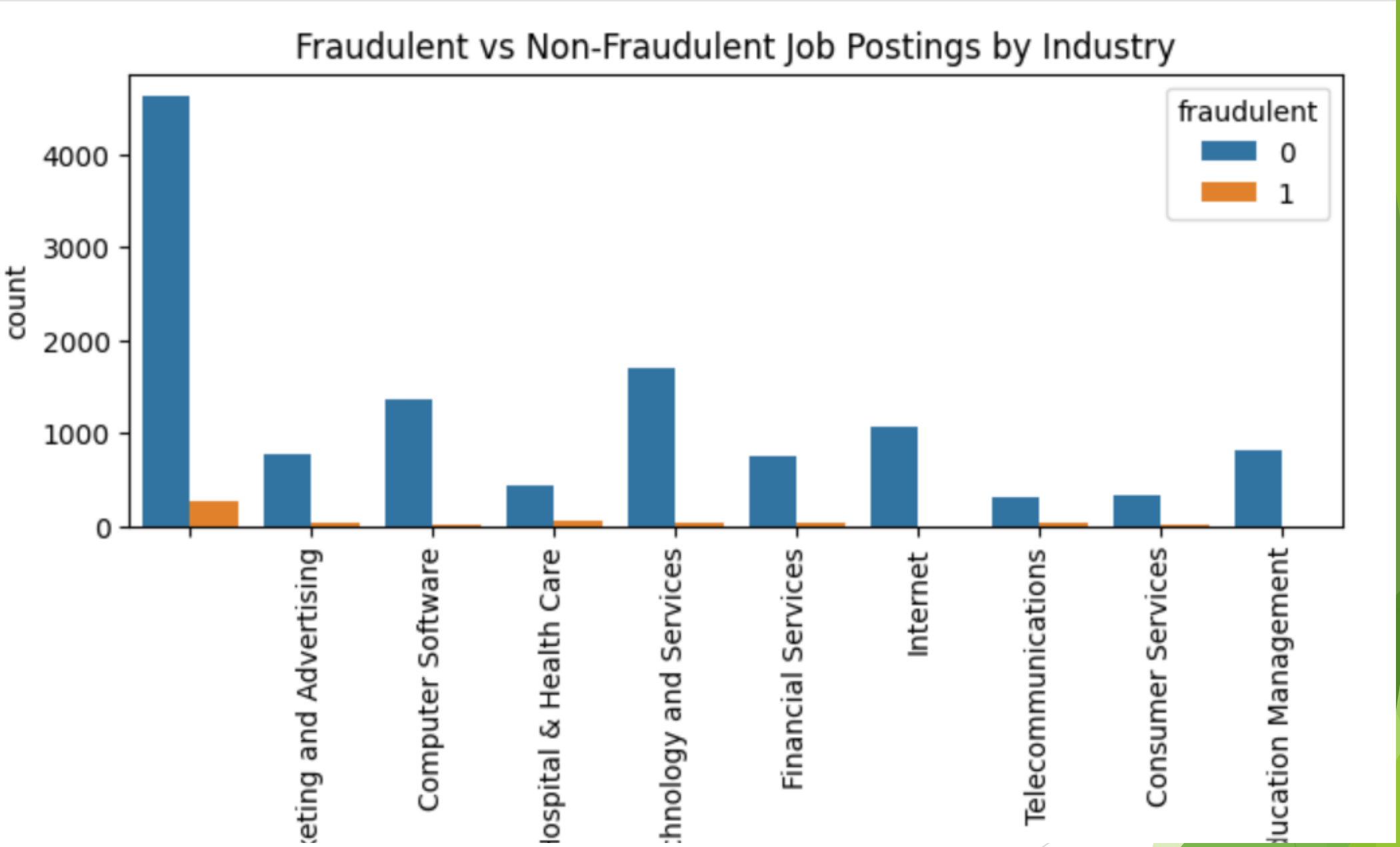
Data Exploration and Preparation

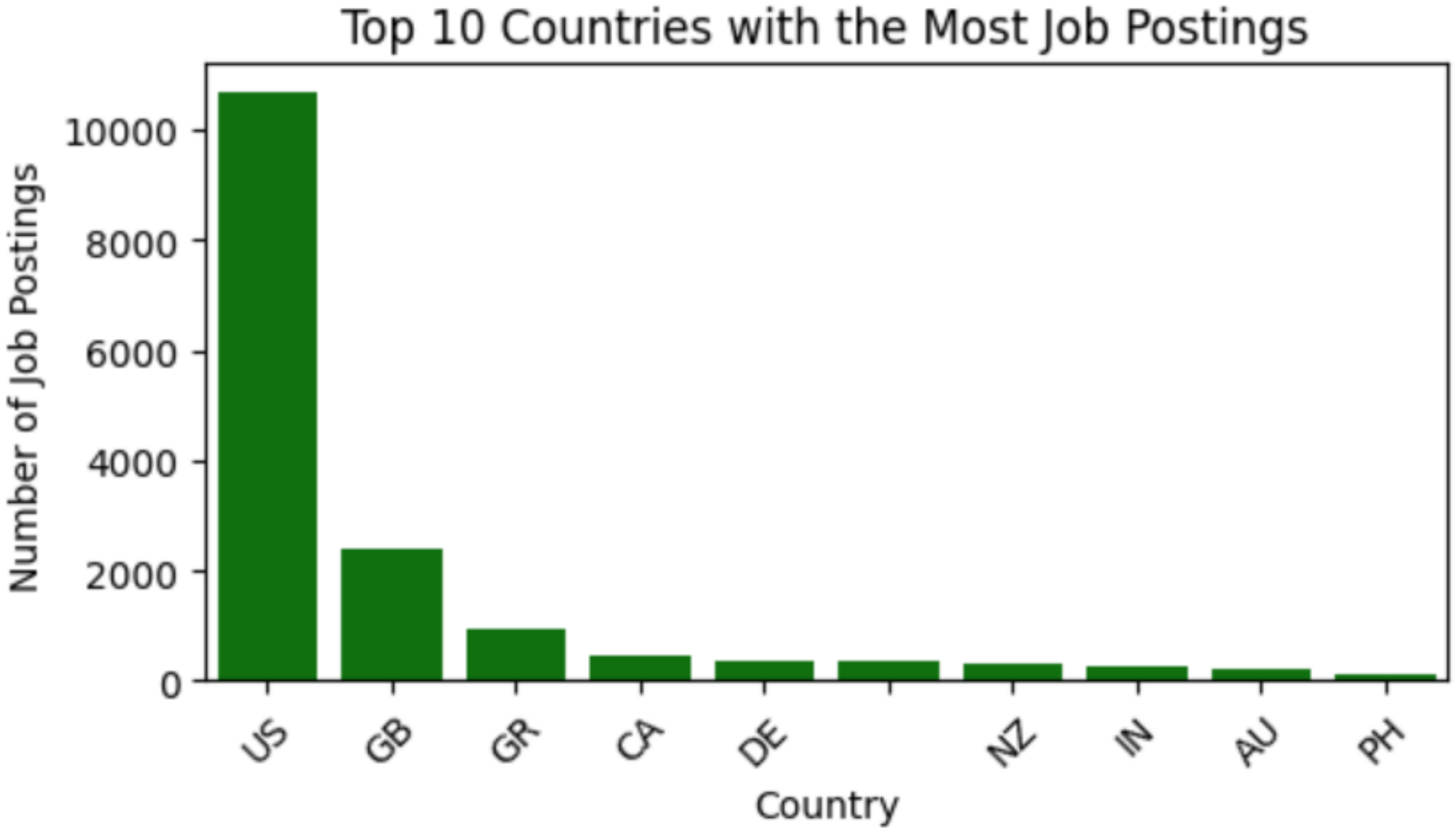
- **Data Structure**
- **Dataset Size**: 17,880 rows and 18 columns.
- **Target Variable**: Fraudulent, binary indicating fake (1) vs real (0).
- **Missing Values Handling**
- **Columns Evaluated**: Significant missing data in salary_range, department.
- **Strategy**: Imputation via median for numerical and mode for categorical data.
- **Data Cleaning**
- **Unnecessary Columns**: Dropped including job_id, telecommuting, and more to focus on relevant features.
- **Preprocessing**: Text cleaned with stemming, tokenization, and stopword removal.


Percentage of Missing Values per Column
Percentage of Missing Values
100
80
60
40
20
0
job_id
title
location
department
salary_range
ompany_profile
description
requirements
benefits
telecommuting
company_logo
has_questions
ployment_type
ed_experience
ired_education
industry
function
fraudulent

Fraudulent vs Real Job postings Countplot




Fraudulent vs Non-Fraudulent Job Postings by Industry
fraudulent
0
1
count
4000
3000
2000
1000
0
Computer Software
Financial Services
Internet
Telecommunications
Consumer Services

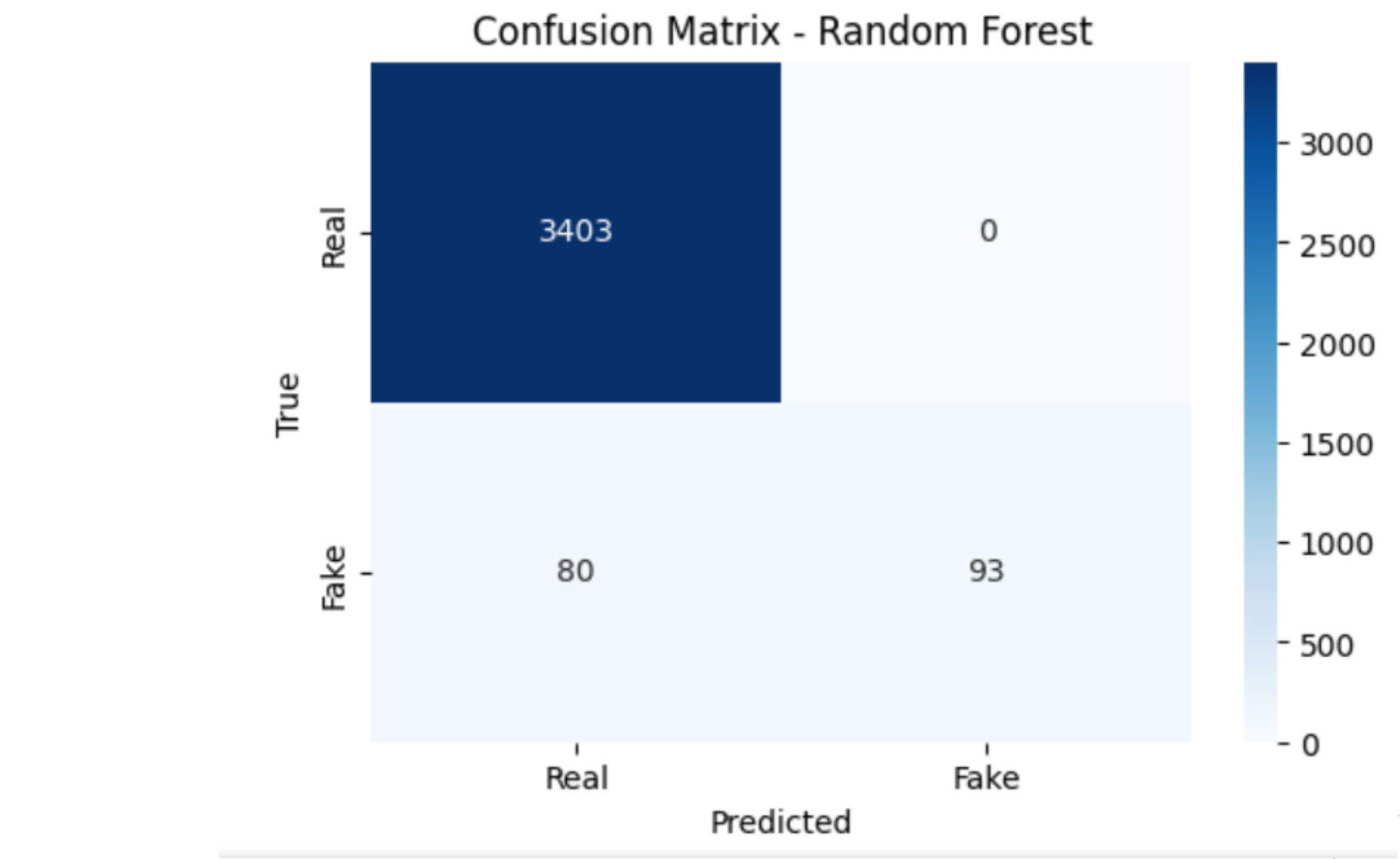

Top 10 Countries with the Most Job Postings
Number of Job Postings
10000
8000
6000
4000
2000
0
US
GB
GR
CA
DE
NZ
IN
AU
PH
Country

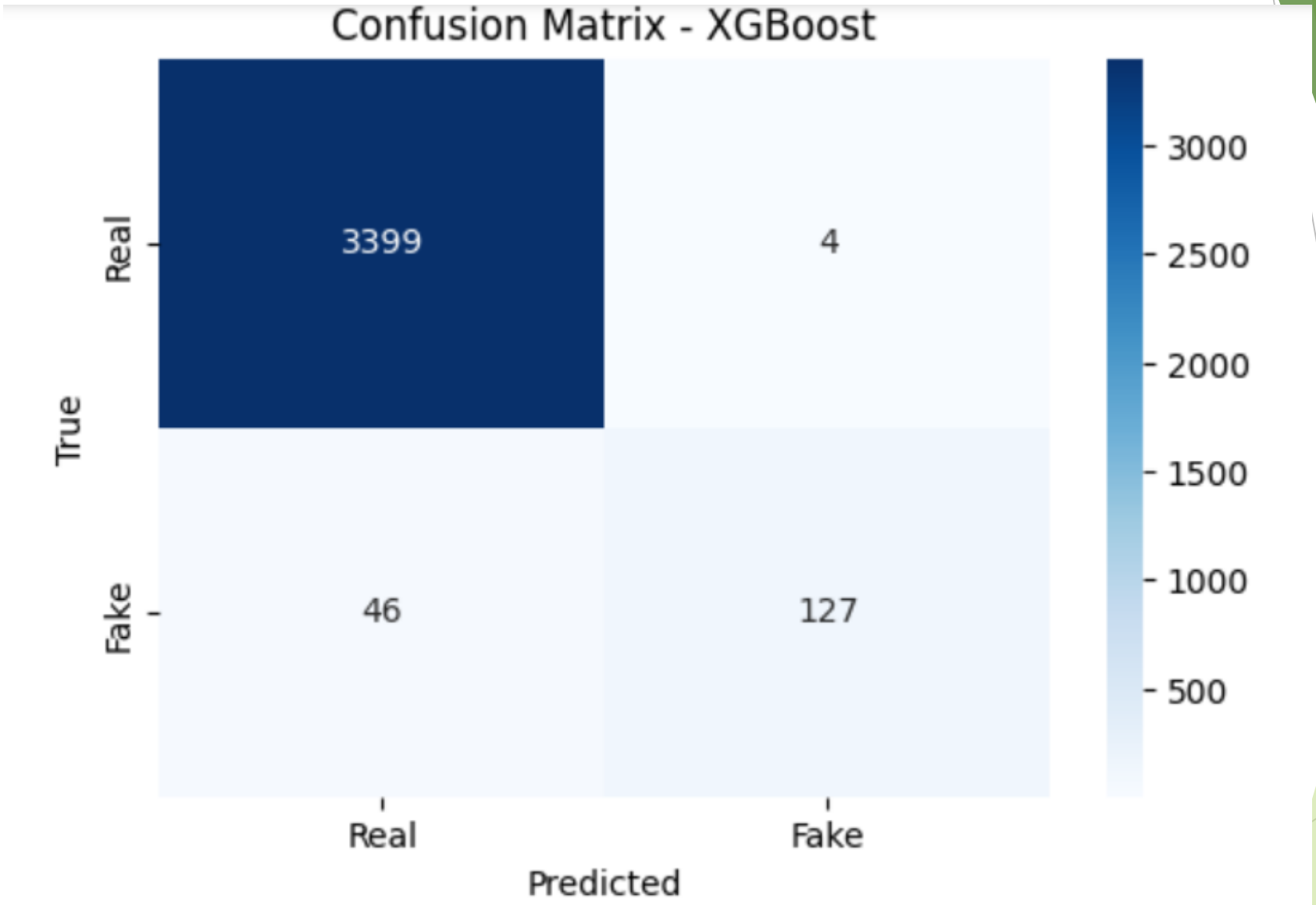
Top 10 Most Common Job Titles

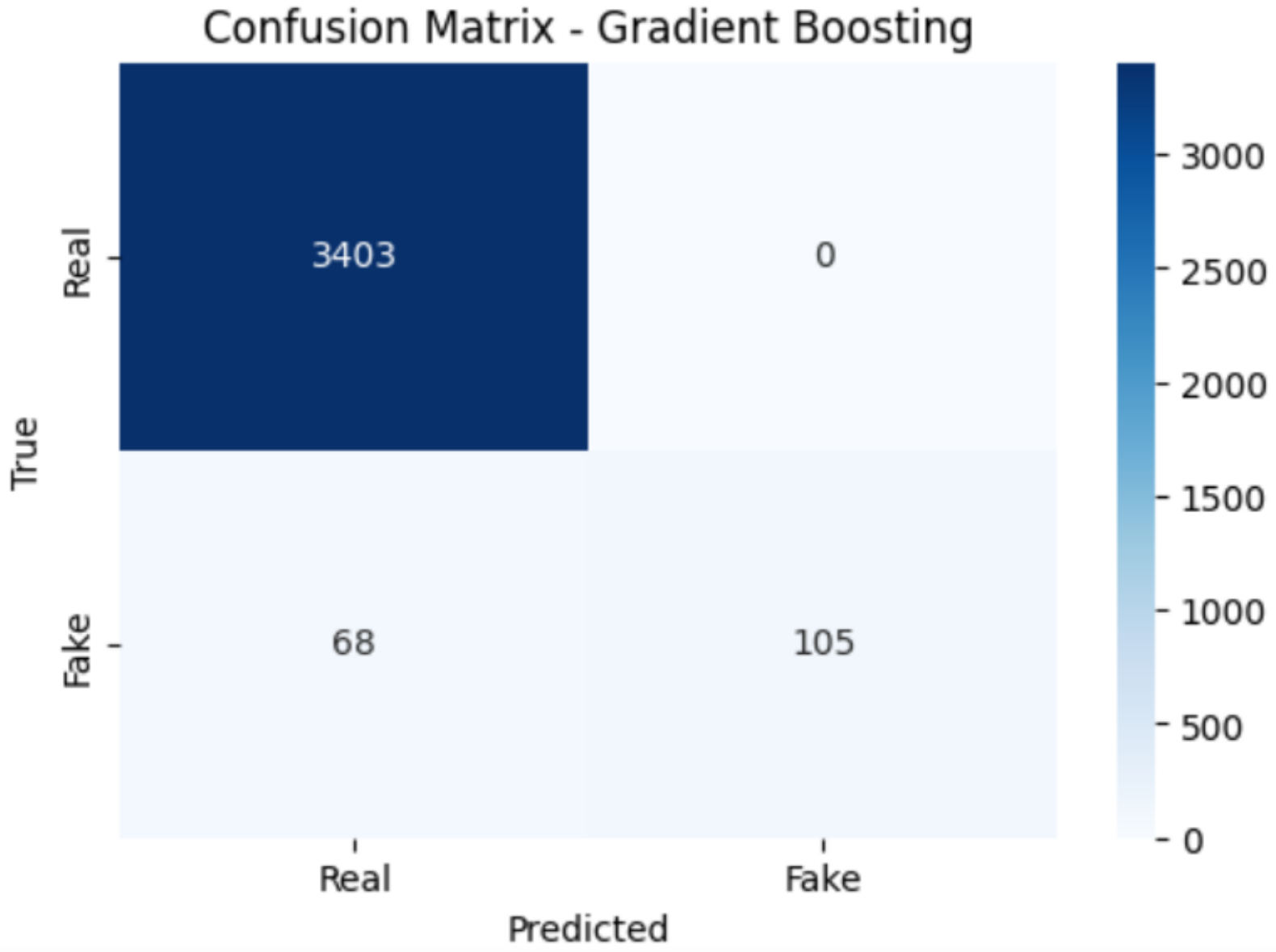
Machine Learning Techniques

2. **K-Nearest Neighbors (KNN)**
    - **Performance**: Accuracy of 98%, yet struggles with minority class.
3. **Random Forest with GridSearchCV**
    - **Performance**: Accuracy of 98%, precision for fake class is high but low recall.
4. **XGBoost**
    - **Performance**: Best overall with accuracy of 99%, high precision and recall.
5. **Gradient Boosting**
    - **Performance**: Accuracy of 98%, similar to Random Forest but slightly better recall.
6. **Naive Bayes**
    - **Performance**: Accuracy of 97.82%, lower recall indicating struggles with false negatives.


Confusion Matrix - KNN
3378
25
47
126
Real
Fake
True
Real
Fake
Predicted
3000
2500
2000
1500
1000
500

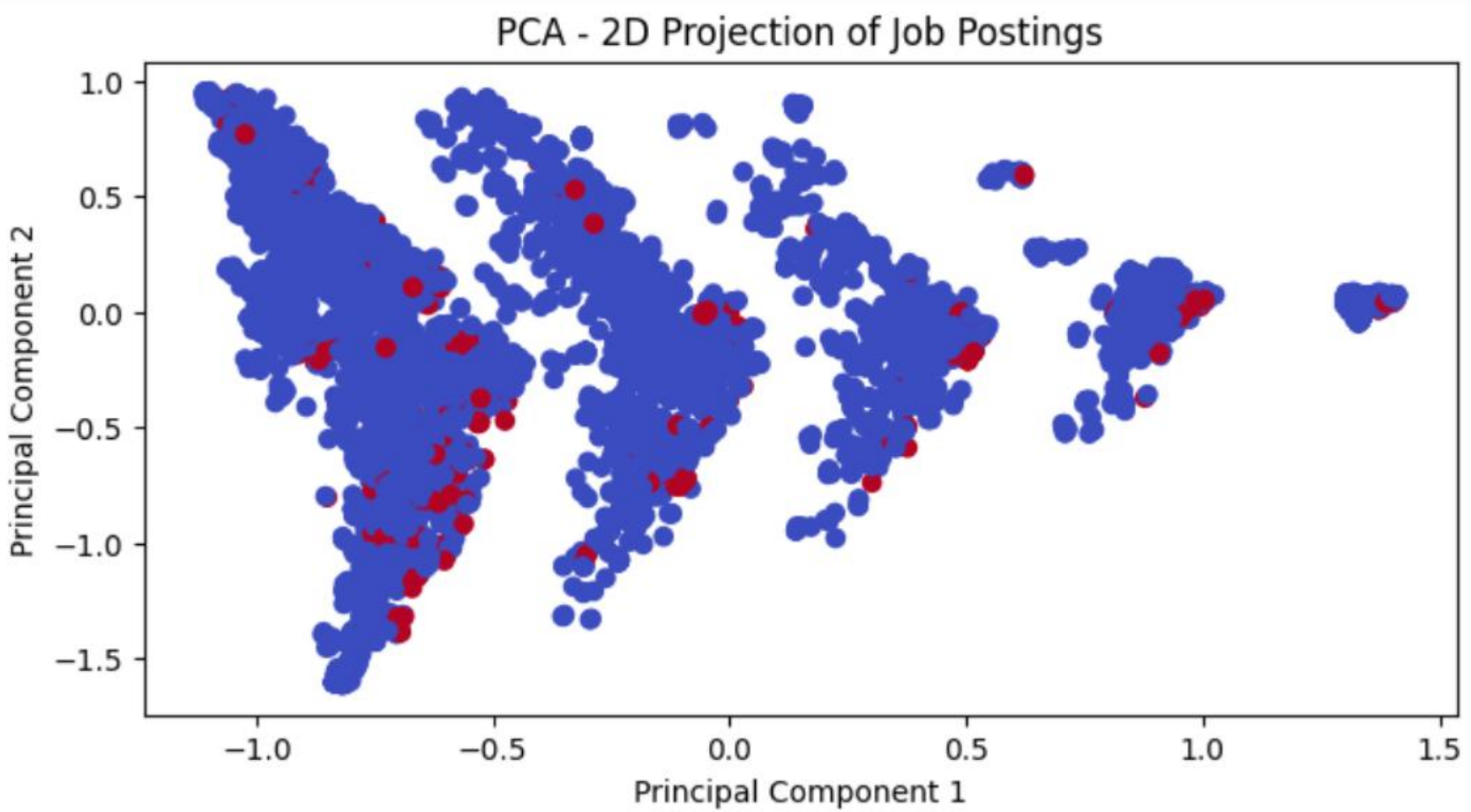

Confusion Matrix - Random Forest
3403
0
80
93
Real
Fake
True
Real
Fake
Predicted
3000
2500
2000
1500
1000
500
0


Confusion Matrix - XGBoost
3399
4
46
127
Real
Fake
True
Real
Fake
Predicted
3000
2500
2000
1500
1000
500


Confusion Matrix - Gradient Boosting
Real
Fake
True
3403
0
68
105
Real
Fake
Predicted
3000
2500
2000
1500
1000
500
0

Unsupervised Learning with PCA

- **PCA Projection**: Data reduced to 2 dimensions helps visualize and simplify the model.



Final Report: Summary

- **Data Analysis**: Robust feature evaluation, cleaning, and transformation led to a high-performing model.
- **Business Value**: Enhances platform credibility, protects job seekers from fraud, and provides actionable insight.