Glassdoor’s Fraud Detection Analysis

Maureen

Wambugu

OUTLINE

- Introduction
- Objective
- Data & Methods
- Results
- Recommendation & Further Analysis
- Questions & End

INTRODUCTION

Overview

- Growing concern over fraudulent job postings on Glassdoor.
- Impact on user trust, platform reputation, and potential legal ramifications.

OBJECTIVE

Develop a reliable classification model to detect and remove fraudulent job postings.

DATA

- Data from Kaggle
- 18K job descriptions including both real and fake job postings.
- Mix of textual and meta-information.
- Features include: Title, Location, Industry, Employment Type, Company Profile, Description, Requirements, Benefits, etc.
- Target Variable: Fraudulent (binary classification).

METHODS

Exploratory Data Analysis

- ❑ Handling Missing Values
- ❑ Categorical feature conversion
- ❑ Feature Correlation with the target

Modeling Approach

- ❑ Logistic Regression
- ❑ Decision Tree
- ❑ Random Forest
- ❑ Gradient Boosting

Evaluation Metrics

- ❑ Precision
- ❑ Recall
- ❑ F1-Score
- ❑ ROC-AUC

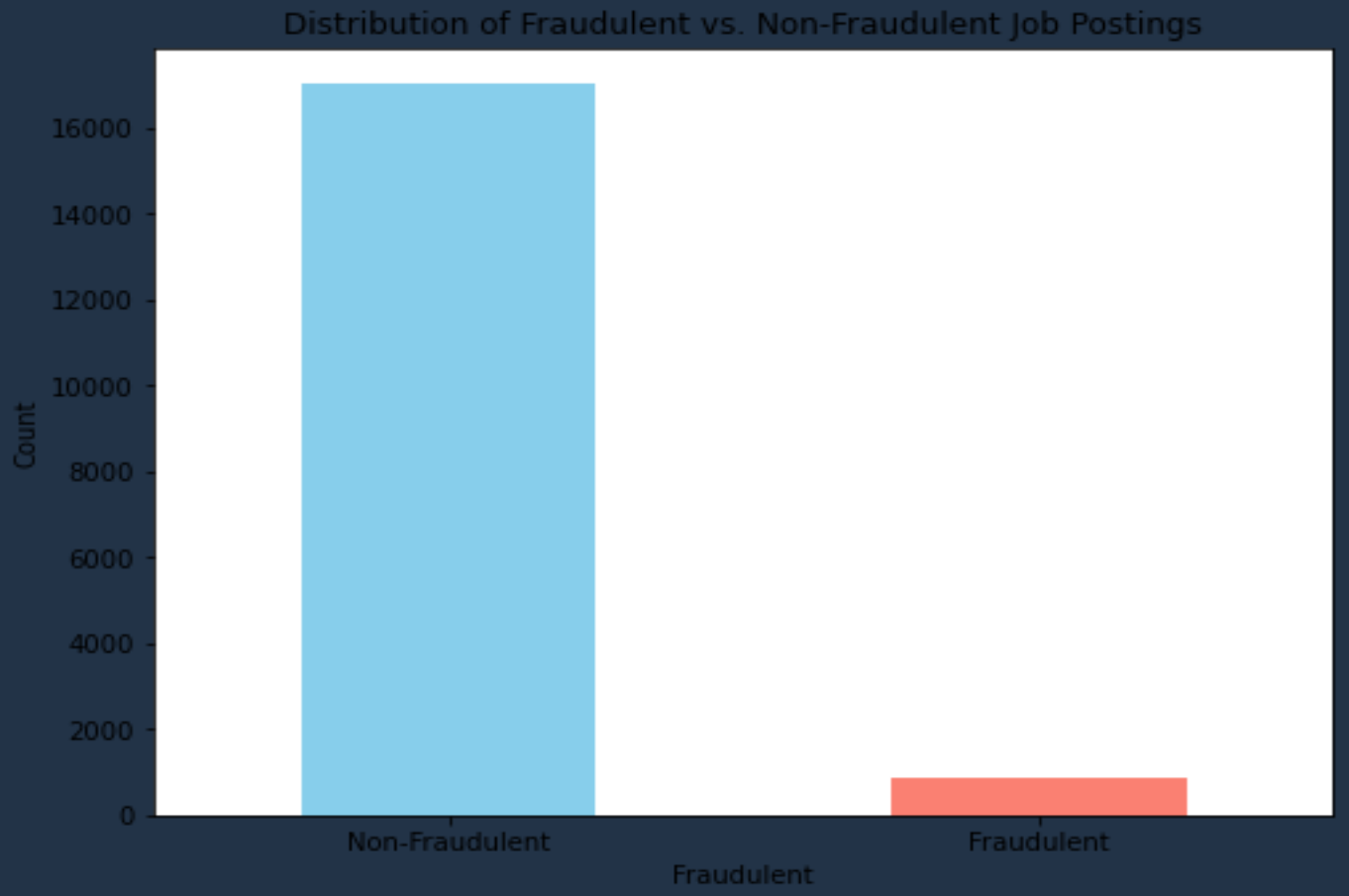
Model Performance Comparison

- ❑ Comparison of model accuracy
- ❑ Best Model Selection

RESULTS

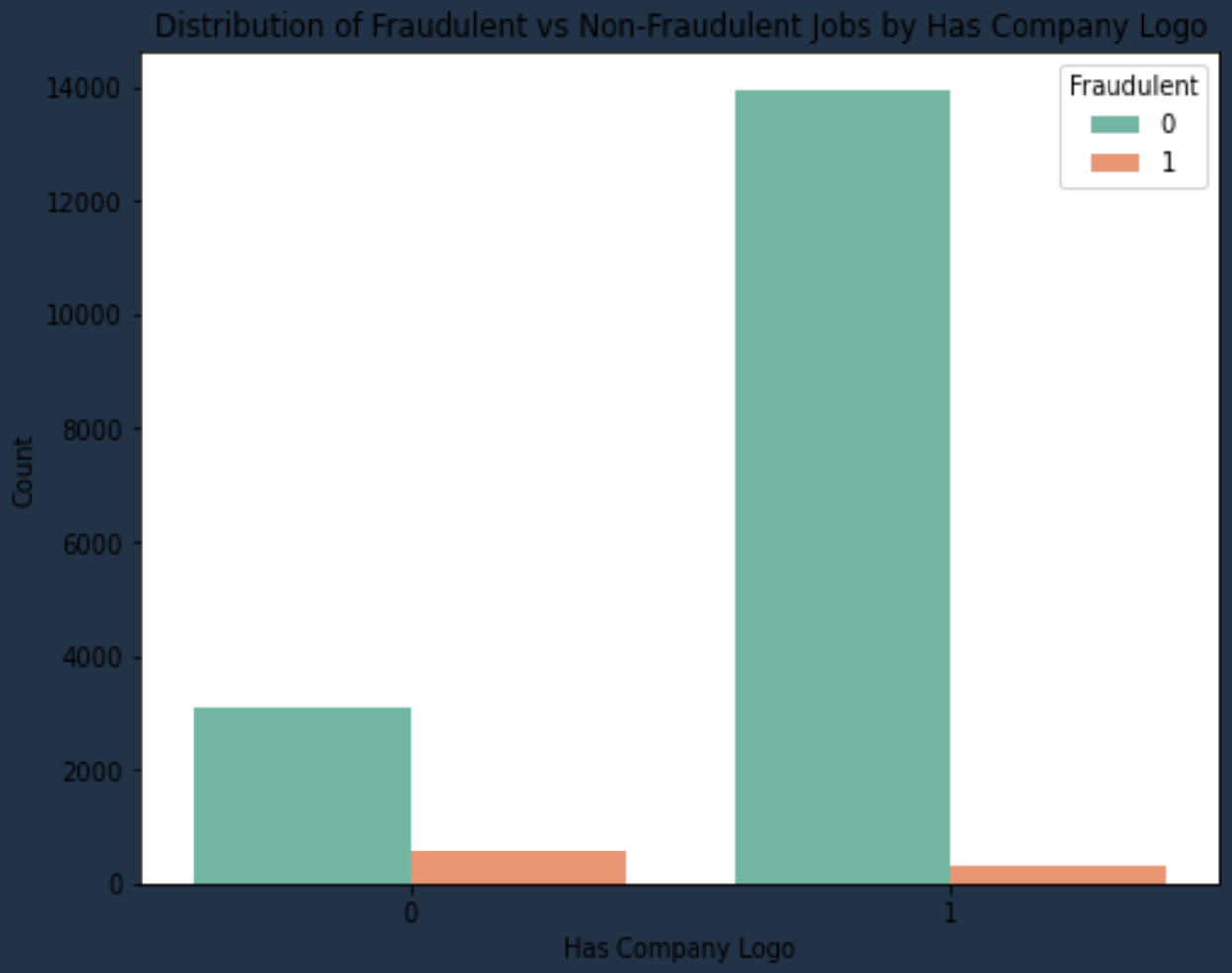
Fraudulent & Non-Fraudulent Distribution

Non-Fraudulent job postings were more as compared to Fraudulent job postings



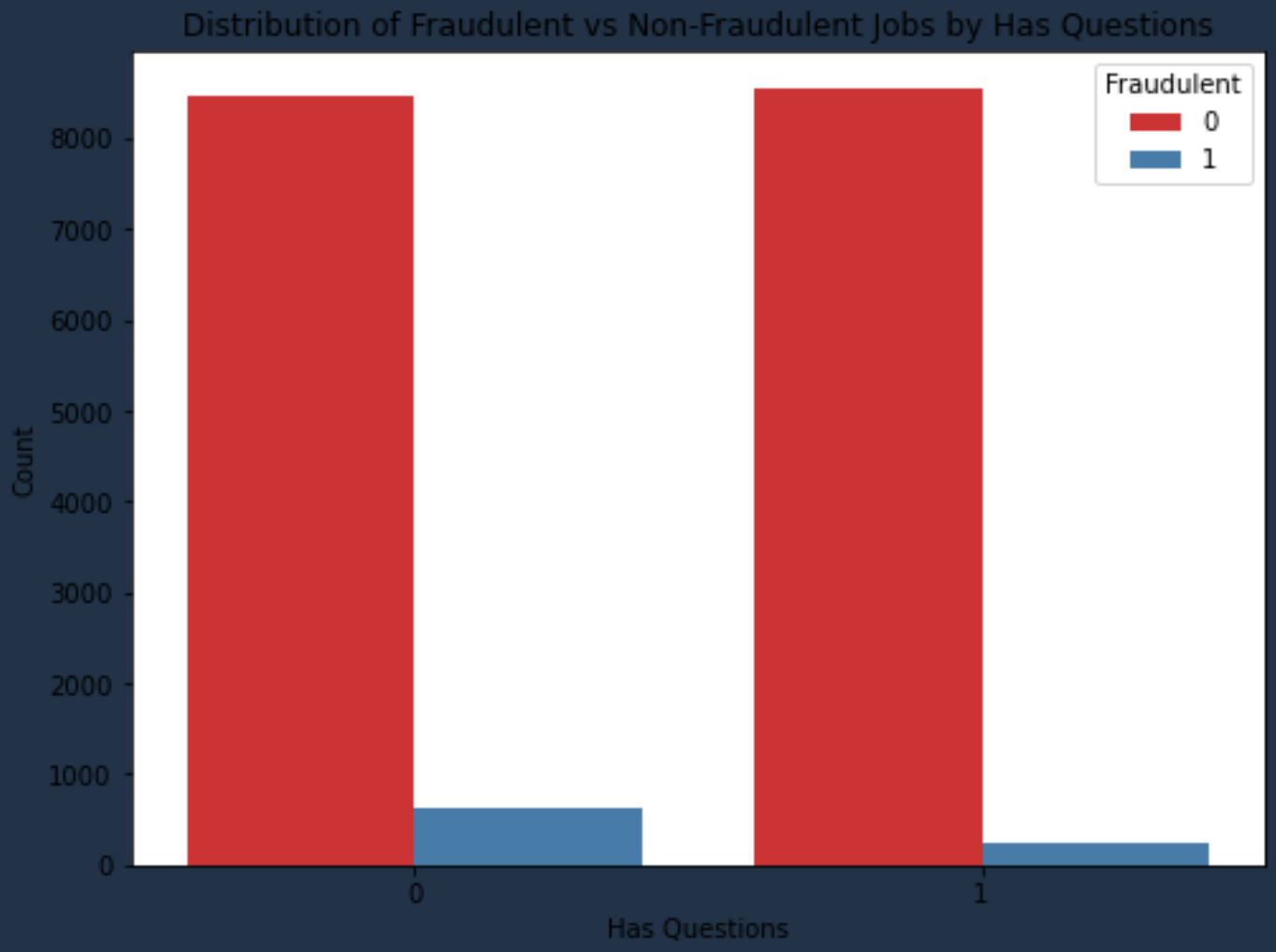
Distribution based on Company logo feature

Job Postings with Company logo have more Non-Fraudulent count and less Fraudulent count as compared to job postings without company logo



Distribution based on Job questions

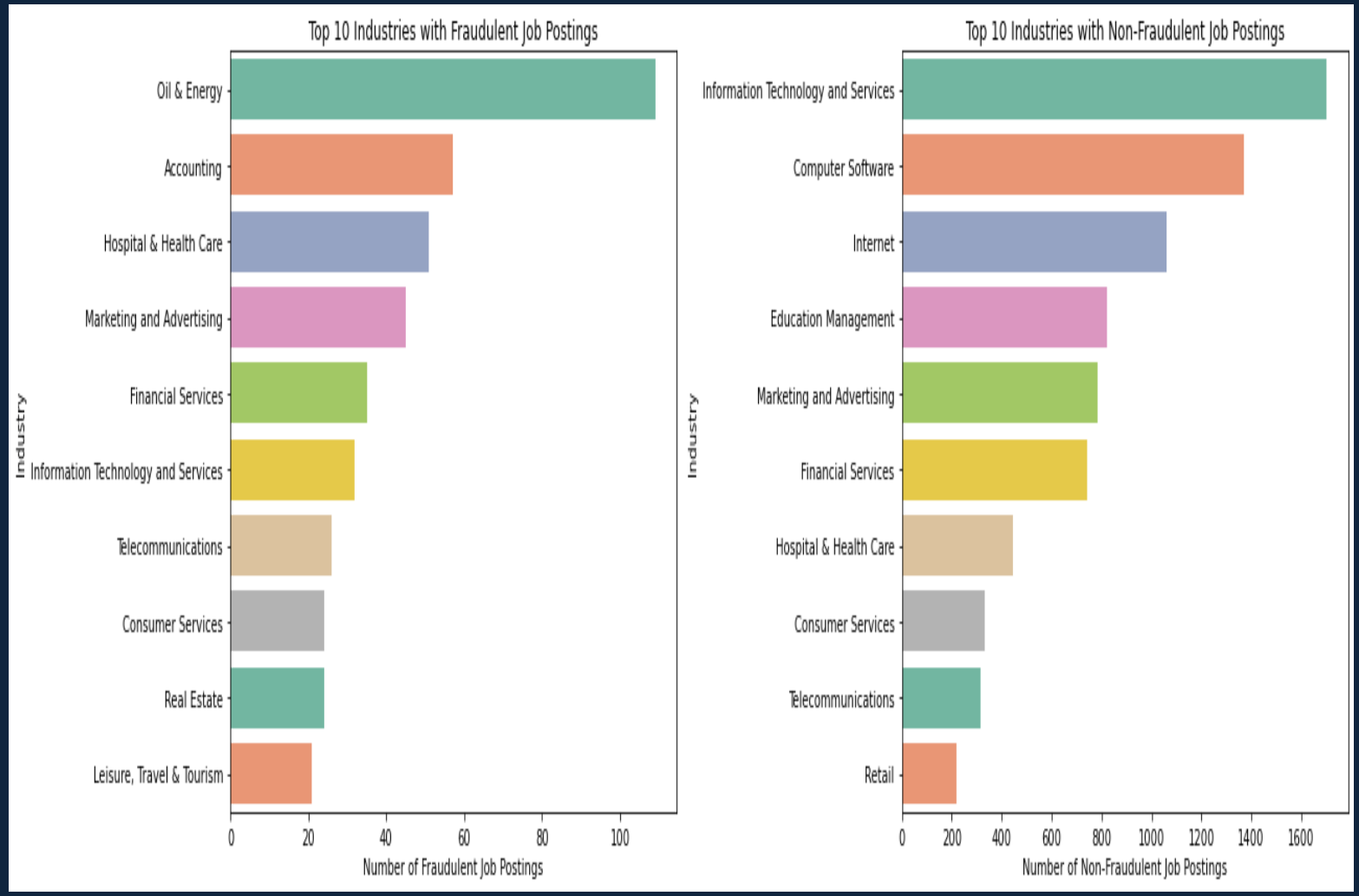
Job Postings with job questions have more Non-Fraudulent count and less Fraudulent count as compared to job postings without job questions



Distribution based on Job Industries

Top 5 Industries with Fraudulent job postings

- Oil & Energy
- Accounting
- Hospital & Health Care
- Marketing & Advertising
- Financial Services



RESULTS

Model Development and Evaluation

- **Logistic Regression:** Good baseline with an accuracy of 95% and a ROC-AUC score of 0.9067.
- **Decision Tree:** Improved accuracy (98%) and precision but slightly lower recall for fraudulent jobs, with a ROC-AUC score of 0.8521.
- **Random Forest:** Best performing model with 98% accuracy and a high ROC-AUC score of 0.9777, excelling in precision and recall.
- **Gradient Boosting:** Strong performance with 98% accuracy and a ROC-AUC score of 0.9432, but slightly lower recall compared to Random Forest.

Model Selection

Random Forest Model

- Best performing model
- Achieved the best balance between precision, recall, and overall accuracy.
- Successfully minimized false negatives, crucial for protecting users from scams.

RECOMMENDATIONS

Enhance User Awareness

- ❑Informing users about common signs of fraudulent job postings

Industry-Specific Monitoring

- ❑Glassdoor should implement stricter verification protocols for job listings in job sectors with high fraudulent count.

Implement Random Forest Model

- ❑Automate the detection and removal of fraudulent postings.

FURTHER STUDIES

- ❑ Exploration of Additional Features
- ❑ Textual Analysis Enhancement

Use of advanced natural language processing (NLP) techniques, such as sentiment analysis or deep learning models

- ❑ Real-Time Fraud Detection

Explore real-time detection capabilities to enhance user safety.

QUESTIONS




THANK YOU!

Name : Maureen Wambugu

GitHub: @Mau-Wambugu

LinkedIn: https://www.linkedin.com/in/maureen-wambugu-02596924b