BT4012 Fraud Analytics

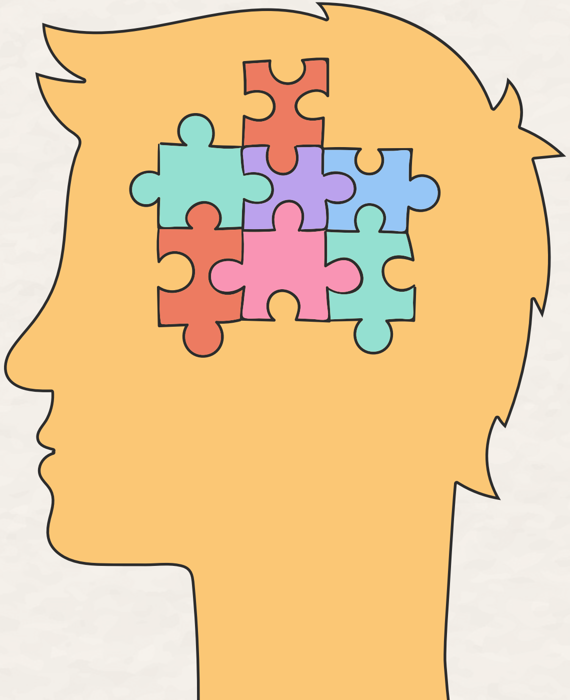
Group 26

A0236495U Lim Zhen Yong
A0234935Y Lam Wen Jett
A0233839W Ng Han Leong, Jordan
A0236437B Lo Zhi Hao

Table of contents

01

Introduction



Introduction to the Problem of Online Job Fraud

Online job portals have become crucial in **connecting job seekers** with potential employers

In the first quarter of **2022**, the US alone recorded over **20,700** cases of job-related frauds, **a third** of which resulted in monetary losses.

Platforms that effectively filter out these scams can gain a **competitive edge**, **retain more users**, and potentially **increase revenue**.

The Problem - Impact on Stakeholders

Job Seekers:
Facing financial and emotional distress

Employers:
Inefficient recruitment process
Disrupts the job market

Job Platforms:
tarnish the reputation of online job platforms.

Our Aim

Enhance the reliability of **online job markets** by developing a **fraud detection system.**

Introduction to Dataset

Name: Employment Scam Aegean Dataset (EMSCAD)

Period: Between 2012 and 2014

Source: real-life job ads posted by Workable, software-as-a-service that provides applicant tracking system and recruitment software

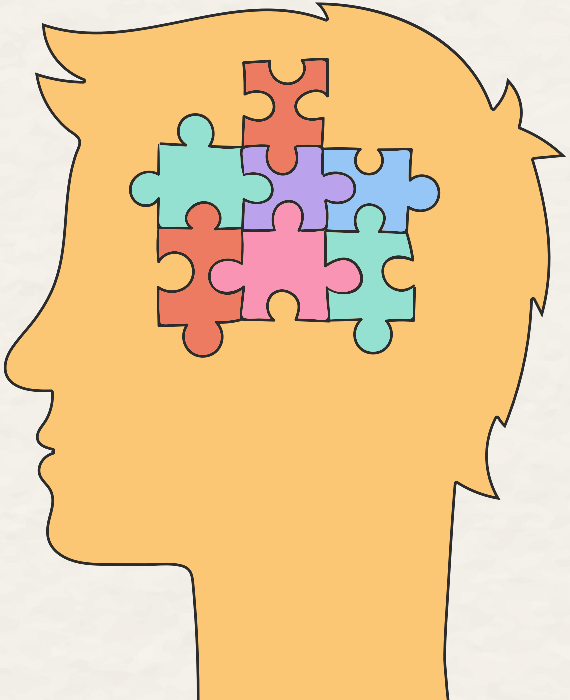
Contributors: All the entries were manually annotated by specialized Workable employees.

Curators: Vidros et al., University of the Aegean

Criteria of Fraud: based on client's suspicious activity on the system, false contact or company information, candidate complaints and periodic detail analysis of the clients.

02

Exploratory Data Analysis



Overview

Univariate Analysis

Bivariate Analysis

Text Data Analysis

Univariate Analysis

Dataset Overview - Null Values Detection - Fraudulent Class Distribution - Feature Distribution (Industry, Department, Location)

Dataset Overview

Dataset: Employment Scam Aegean Dataset (EMSCAD)

Dimensions: 17880 rows and 18 columns (including target variable)

Target Variable: fraudulent (Class 0 - Non-Fraudulent, Class 1 - Fraudulent)

Features: 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

```
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   job_id               17880 non-null  int64 
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15184 non-null  object
 8   benefits             10668 non-null  object
 9   telecommuting        17880 non-null  int64 
 10  has_company_logo     17880 non-null  int64 
 11  has_questions        17880 non-null  int64 
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64 
dtypes: int64(5), object(13)
memory usage: 2.5+ MB
```

Features Overview

telecommuting, has_company_logo, has_questions

Binary Features

location, department, salary_range, employment_type, required_experience, required_education, industry, function

Multiclass Categorical Features

company_profile, description, requirements, benefits

Text Features

Null Values Detection

Overall, out of the 17 features columns in the dataset, ***12*** features have ***missing values***.

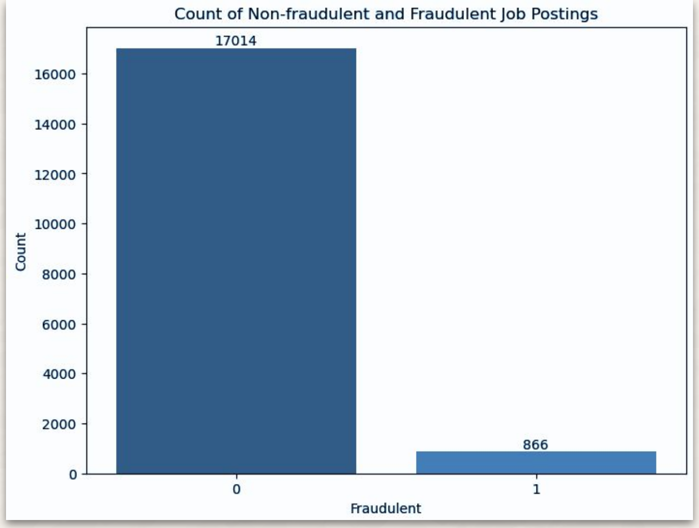
The features that have missing values includes (from least to most):

1. description
2. location
3. requirements
4. company_profile
5. employment_type
6. industry
7. function
8. required_experience
9. benefits
10. required_education
11. department
12. salary_range

job_id	0
description_length	0
company_profile_length	0
fraudulent	0
has_questions	0
requirements_length	0
telecommuting	0
has_company_logo	0
title	0
benefits_length	0
description	1
location	346
requirements	2696
company_profile	3308
employment_type	3471
industry	4903
function	6455
required_experience	7050
benefits	7212
required_education	8105
department	11547
salary_range	15012
dtype: int64	

Fraudulent Class Distribution

We noticed significant ***class imbalance*** in the Fraudulent target class, with ***17,014 instances of non-fraudulent*** job postings, and only ***866 instances of fraudulent*** jobs.

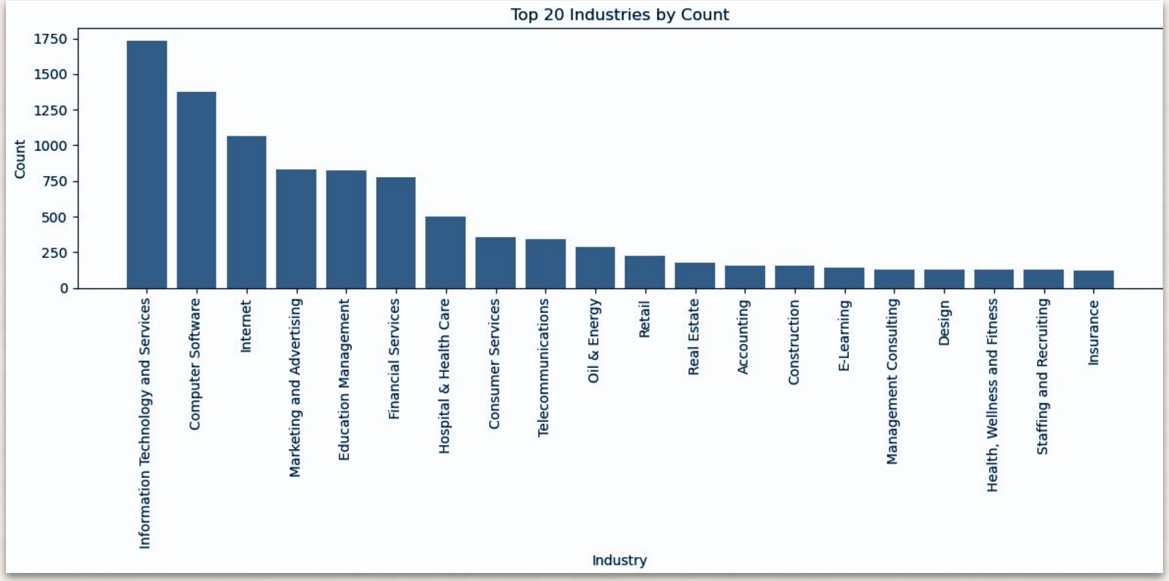


95% Non-Fraudulent *VS* **5%** Fraudulent

Feature Distribution – Industry

There are ***133 unique industries*** that are observed in total. Top 3 industries are ***IT and Services***, ***Computer Software*** and ***Internet***.

However, ***6 industries*** has only ***1 observation***, while ***42 industries*** have ***less than 10 observations***.

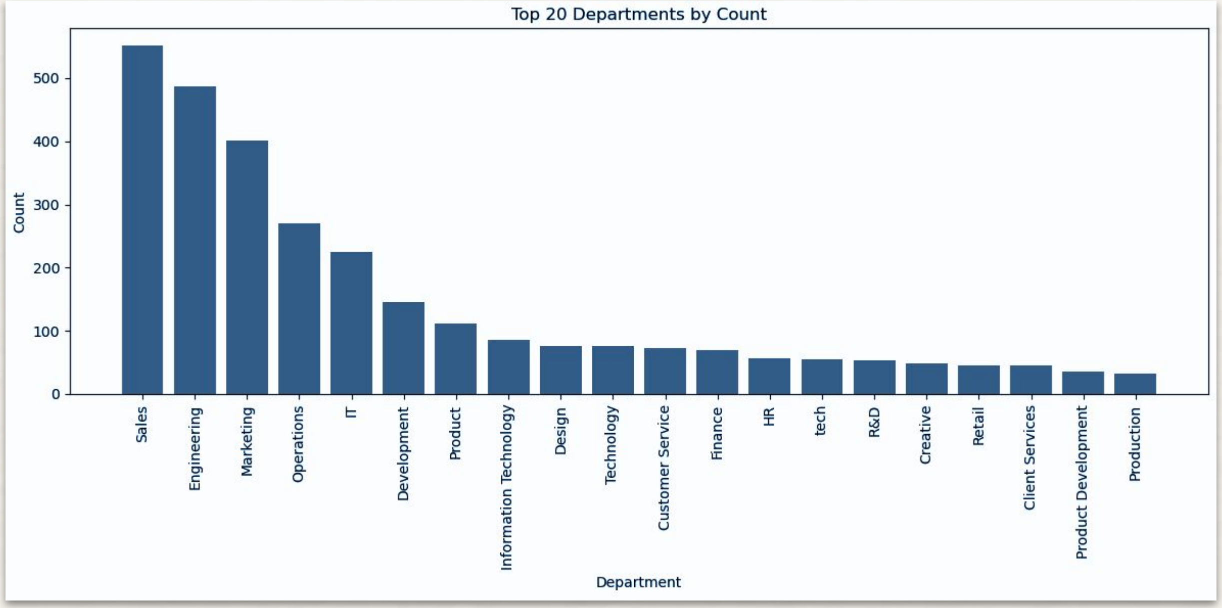


Industry	Count
Alternative Dispute Resolution	1
Shipbuilding	1
Sporting Goods	1
Museums and Institutions	1
Wine and Spirits	1
Ranching	1

Feature Distribution – Departments

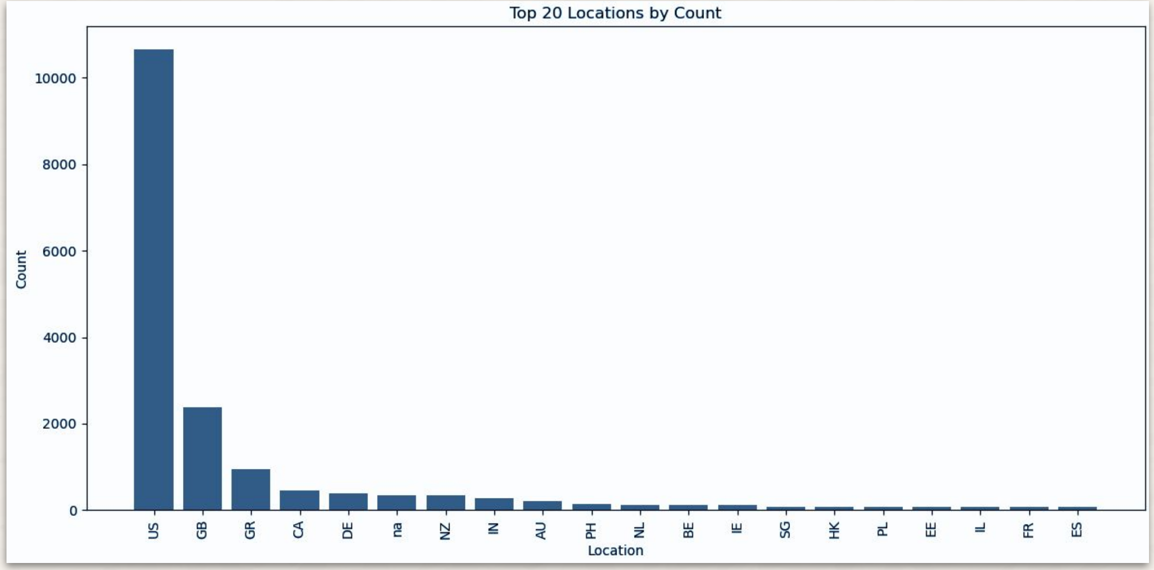
There are ***1,337 unique departments*** that are observed in total. Top 3 departments are ***Sales***, ***Engineering*** and ***Marketing***.

However, ***815*** departments has only ***1 observation***, while ***1256*** departments have ***less than 10 observations***.



Feature Distribution - Location

In terms of location, we processed it into country and cities (details will be shared later!). There are ***91 unique countries*** that are observed. Top 3 locations of postings are ***United States***, ***Great Britain*** and ***Greece***.



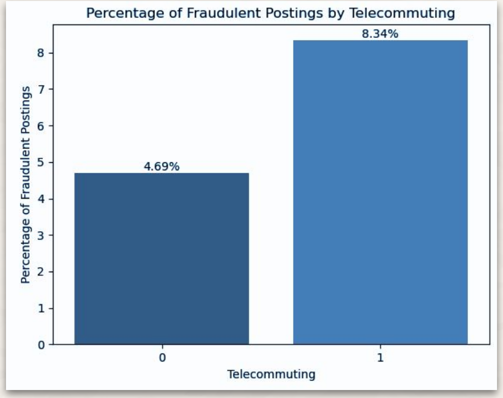
However, ***14*** countries has only ***1 observation***, while ***38*** countries have ***less than 10 observations***.

Bivariate Analysis

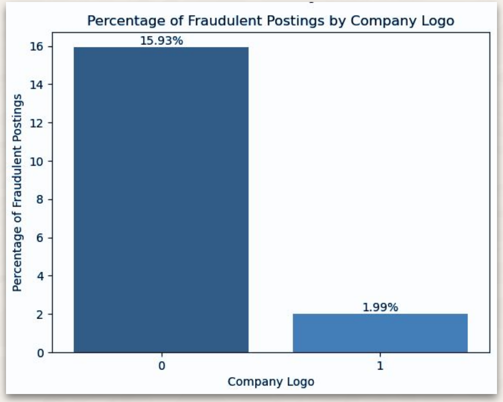
Fraudulent Postings by Features (Telecommuting, Have Company Logo, Have Questions, Employment Type, Function, Required_education, Required_experience)

Binary Features *vs* Fraudulent

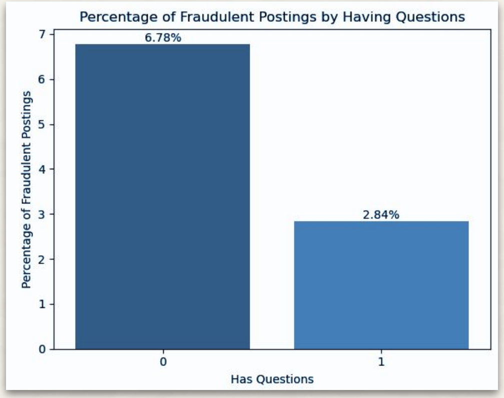
Telecommuting



Have Company Logo



Has Questions

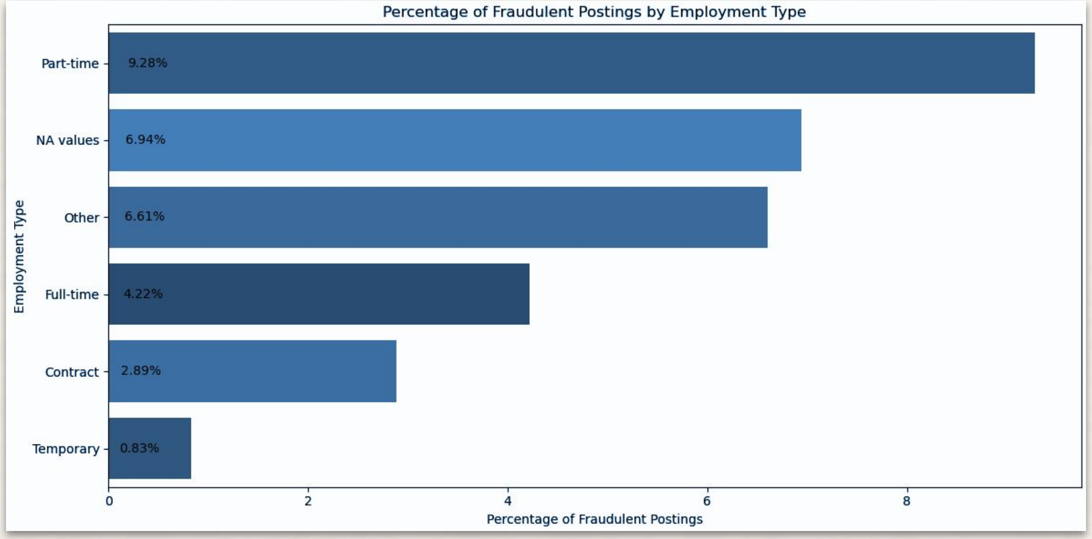


We noticed significant differences in terms of chances of fraudulent job postings across the binary features. Job that ***allows telecommuting***, ***does not have screening questions***, and ***posted by companies which does not provide company logo*** are more likely to be fraudulent.

Employment Type *vs* Fraudulent

Part-time job postings have a higher chance of being fraudulent.

When ***employment type is not provided (Missing Value)*** or is stated as ***others***, the chances of the job posting being fraudulent is higher as well.

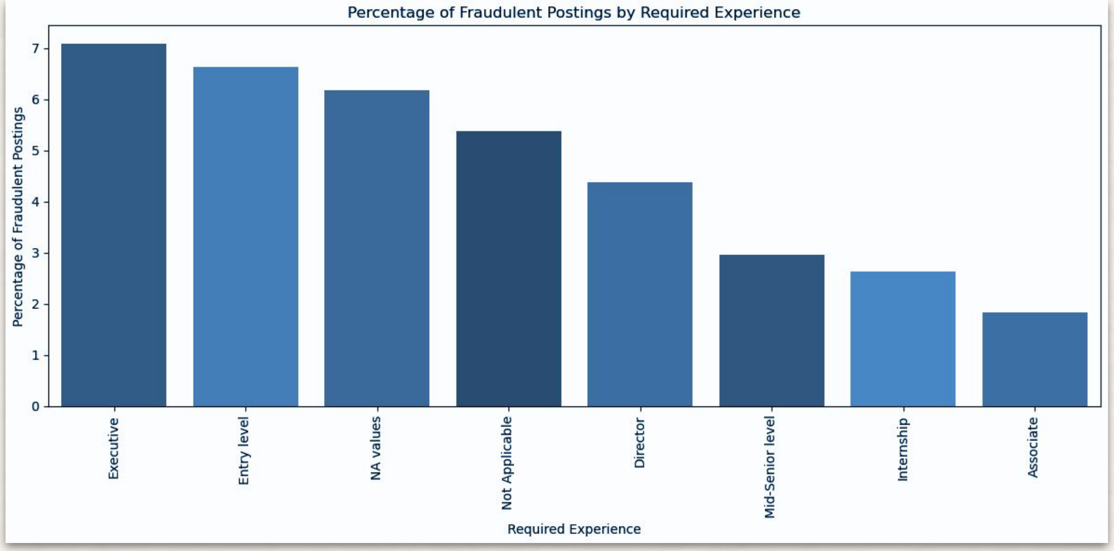


	Occurrences	Percentage of Fraudulent
employment_type		
Part-time	797	0.092848
NA values	3471	0.069432
Other	227	0.066079
Full-time	11620	0.042169
Contract	1524	0.028871
Temporary	241	0.008299

Required Experience *vs* Fraudulent

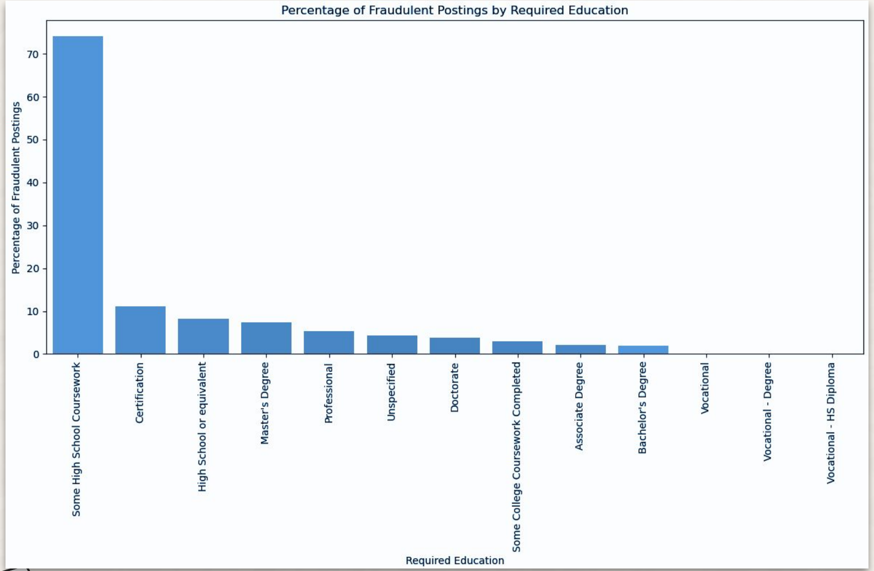
Executive and ***Entry Level*** required experience have a higher chance of being fraudulent as compared to others.

We should also note that there is a higher chance of fraud when required experience are ***not available*** or ***not applicable***.



required_experience	Occurrences	Percentage of Fraudulent
Associate	2297	1.828472
Director	389	4.370180
Entry level	2697	6.637004
Executive	141	7.092199
Internship	381	2.624672
Mid-Senior level	3809	2.966658
NA values	7050	6.170213
Not Applicable	1116	5.376344

Required Education *vs* Fraudulent



required_education	Occurrences	Percentage of Fraudulent
Associate Degree	274	2.189781
Bachelor's Degree	5145	1.943635
Certification	170	11.176471
Doctorate	26	3.846154
High School or equivalent	2080	8.173077
Master's Degree	416	7.451923
NA values	8105	5.564466
Professional	74	5.405405
Some College Coursework Completed	102	2.941176
Some High School Coursework	27	74.074074
Unspecified	1397	4.366500
Vocational	49	0.000000
Vocational - Degree	6	0.000000
Vocational - HS Diploma	9	0.000000

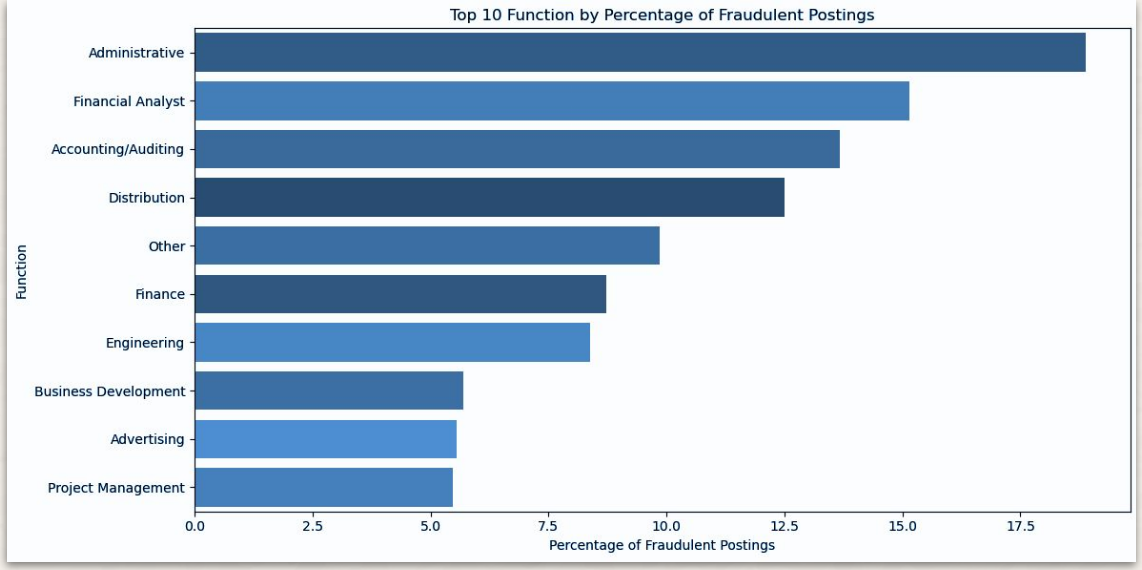
In terms of education, we note that positions that require ***high school*** or ***certification*** education levels are more likely to be fraudulent.

Positions that requires ***master's degree*** also exhibited higher chance of being fraudulent.

Function *vs* Fraudulent

In terms of job functions, we notice that ***Administrative*** functions, ***Financial Analyst*** and ***Accounting/Auditing*** have the highest chance of being fraudulent.

Coupled with the observations from required education and required experience, we do notice that ***most of the fraudulent postings are targeted towards less educated and experienced personnels***.



Text Data Analysis

Company Profile Analysis - Job Requirements Analysis - Job Title Analysis - Job Description Analysis - Job Benefits Analysis

Job Titles *vs* Fraudulent

Fraudulent Job Titles
Data Entry Admin/Clerical Positions - Work From Home
Cruise Staff Wanted *URGENT*
Home Based Payroll Data Entry Clerk Position - Earn $100-$200 Daily
Account Sales Managers $80-$130,000/yr
Payroll Clerk

Non-Fraudulent Job Titles
English Teacher Abroad
Customer Service Associate
Software Engineer
Account Manager
Project Manager

Job titles in fraudulent postings tend to be more ***entry level***, and tend to include details such as ***salary*** and ***special characters***.

Job Details *vs* Fraudulent

Word Cloud for Fraudulent Job Descriptions

Word Cloud for Fraudulent Job Benefits

Word Cloud for Fraudulent Job Requirements

In terms of job descriptions, requirements and benefits, we notice a few subtle difference between fraudulent and non-fraudulent posts.

This includes more frequent use of words such as `***project***`, `***online***` and `***require***`, among the rest.

Thus, from our analysis, we do believe that ***text data are important*** and should be processed as features for our classification model.

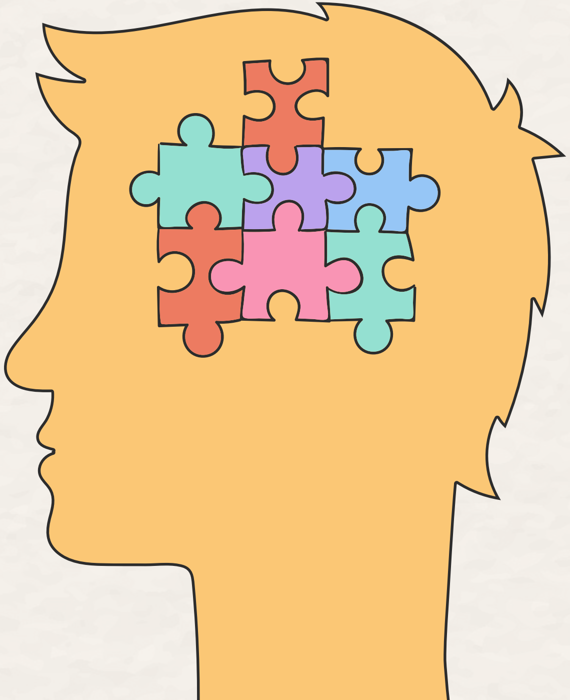
Word Cloud for Non-Fraudulent Job Descriptions

Word Cloud for Non-Fraudulent Job Benefits

Word Cloud for Non-Fraudulent Job Requirements

03

Data Preprocessing



Data Preprocessing Techniques

Handling Missing Values

Feature Engineering

Text Preprocessing

Text Encoding

SMOTE Oversampling

Handling of Missing Values

After performing Exploratory Data Analysis:

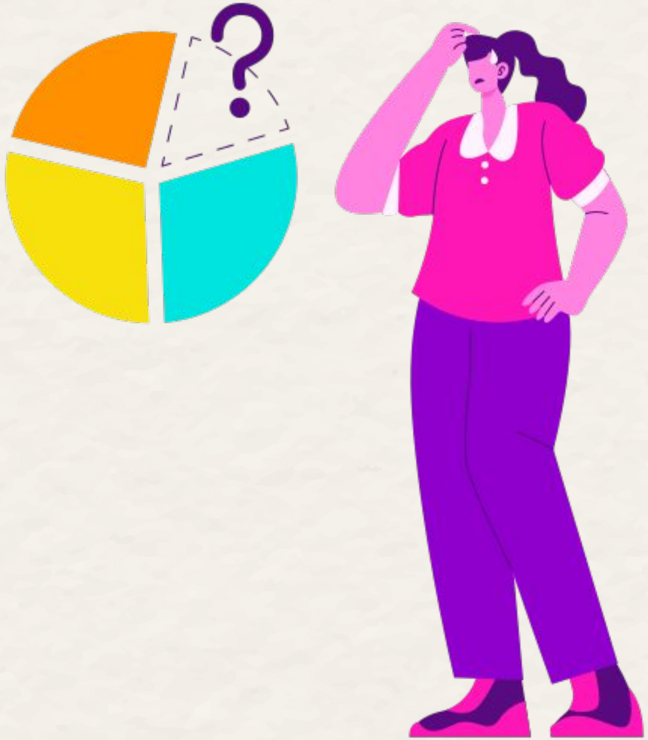
- Categorical → “Unknown”
- Text → “No available data”

Reasoning:

- Lack of context in data collection and job postings details → **Missing Completely At Random (MCAR)**

Application:

- Applied across all columns with NA values for a consistent approach



Feature Engineering

Column-specific feature engineering:

- 'Location' → Split into 'country' and 'city', followed by one hot encoding
- 'Department' → Categorize rare departments as 'others'

Reasoning:

- Ensure **interpretable data** for model
- Mitigate **overfitting & reduce training data dimensionality**

Application:

- Applied to 'Country', 'city' and 'industry' columns as well

Text Preprocessing

From our Exploratory Data Analysis:

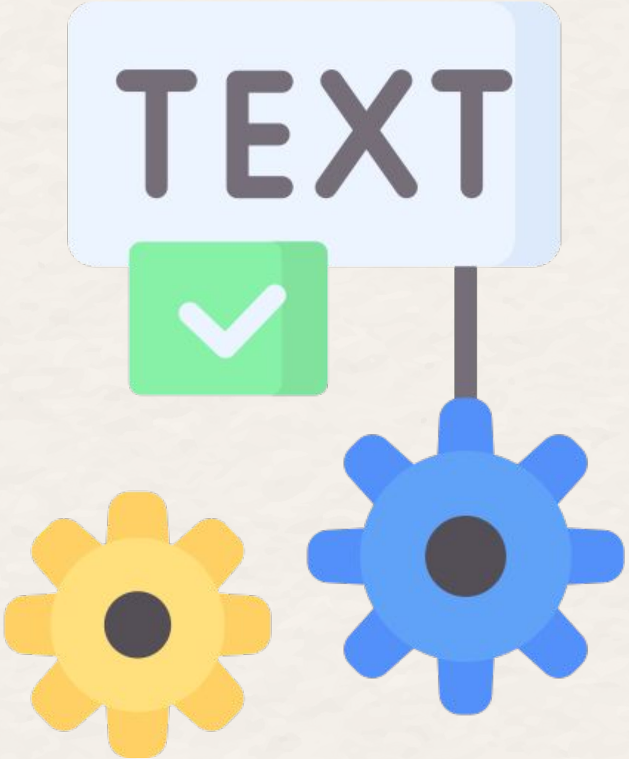
- Significant amount of text data (‘title’, ‘company_profile’, ‘description’, ‘requirements’, ‘benefits’)
- Capture context of text → **high value features**

Text Preprocessing Methods:

- Removing stopwords → reduce noise & enhance relevance of text data
- Tokenization → individual words
- Stemming and Lemmatization → basic form

Application:

- Applied to ‘full_text’ column → contains concatenated text data from key columns



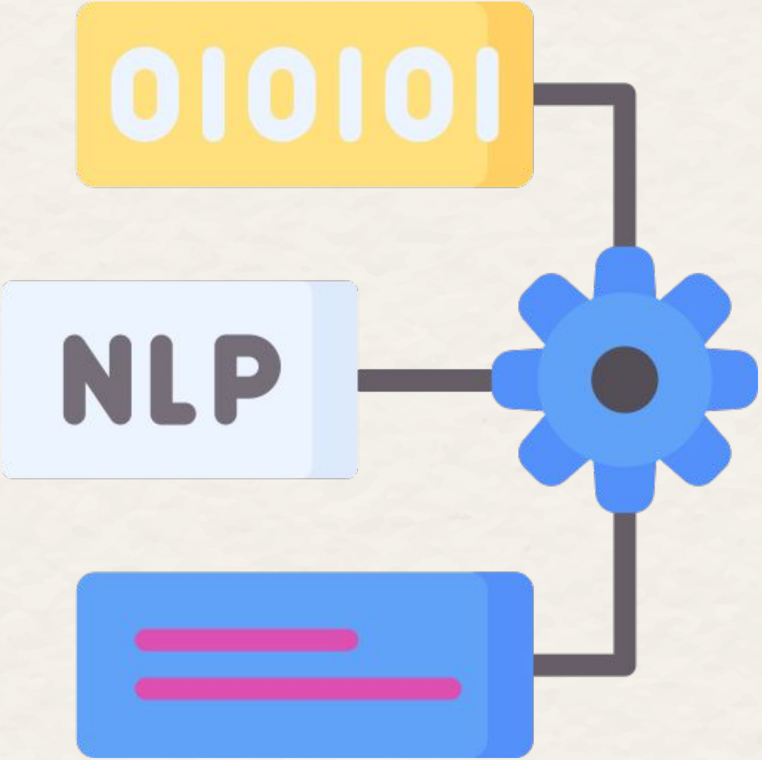
Text Encoding

Transforming text data into machine-readable format:

- Bag of Words (BoW) → statistical embeddings of vocabularies
- Pre Trained Word2Vec Word Embedding → preprocessing method that yield the best features

Reasoning:

- Enhance **significance of features** derived from text data
- Statistical embeddings **capture nuanced relationships** within the text



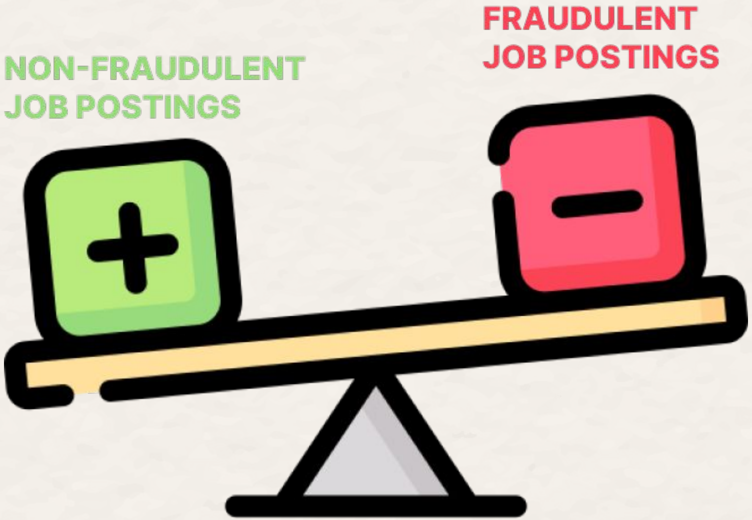
SMOTE Oversampling

Due to an imbalanced dataset:

- Utilize Synthetic Minority Oversampling Technique (SMOTE)

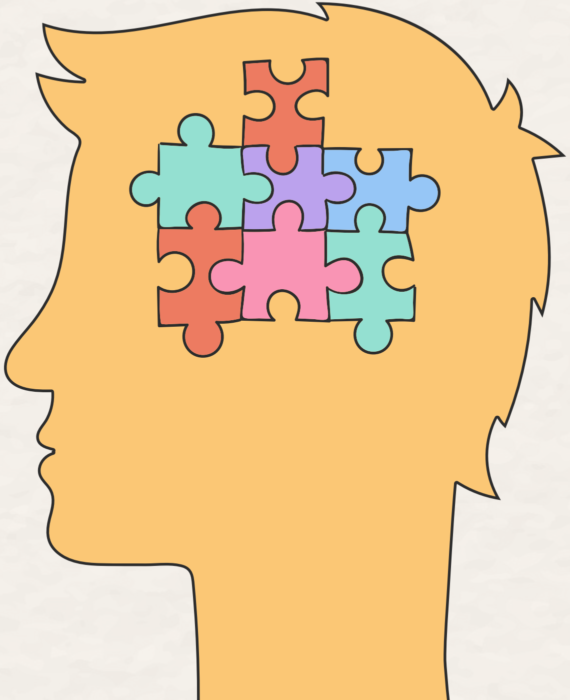
Reasoning:

- Address class imbalances → improve **model performance**
- Counter biases towards majority class → prevent **model skewing**
- Enhance **generalizability** to unseen data



04

Model Evaluation



Metrics for Evaluation

Accuracy

Measures **overall correctness** by calculating the ratio of correctly predicted instances to the total number of instances.

Roc AUC

Illustrates the **trade-off** between **true positive rate** and **false positive rate.**

Precision

Assesses the **accuracy of positive predictions** by calculating the ratio of **true positives** to the sum of **true positives and false positives.**

Recall

Correctly identify **all relevant instances of a class**, calculated as the ratio of **true positives** to the sum of **true positives and false negatives**.

F1 Score

Combines **precision and recall** into a single metric, providing a **balance between false positives and false negatives** in a model's performance.

Best Performers (Before SMOTE)

Model	Accuracy	F1-Score	Precision	Recall	Roc AUC
LSTM (With CountVec)	99%	**87%**	97%	78%	81%
Decision Tree Classifier (With CountVec)	98%	78%	76%	**81%**	90%

Best Performers (After SMOTE)

Model	Accuracy	F1-Score	Precision	Recall	Roc AUC
XGBClassifier (With Word2Vec)	99%	**87%**	93%	82%	99%
KNN Classifier (With TF-IDF)	86%	42%	27% ?	**97%**	95%

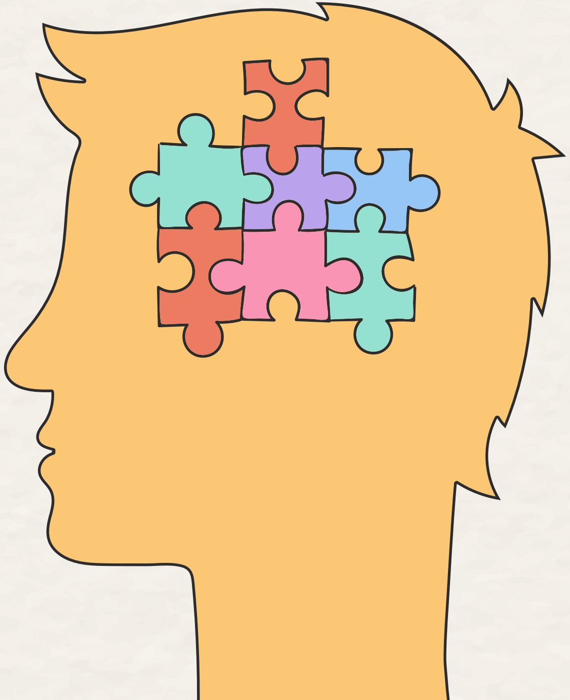
Optimal Model Selected

Model	Accuracy	F1-Score	Precision	Recall	Roc AUC
XGBClassifier (With Word2Vec without SMOTE)	98%	80%	**98%**	**67%**	99%
XGBClassifier (With Word2Vec and SMOTE)	99%	87%	**93%**	**82%**	99%

- The Recall score increased for the model after oversampling.
- Although this comes at the cost of lower Precision, the cost of false negative is much high than false positive.

05

Conclusion & Future Works



Conclusion - Project Achievements

- successfully developed a machine learning-based fraud detection system
- combines data preprocessing, feature engineering, and the implementation of various classifiers to identify fraudulent job postings
- determined the most effective classifier

Integration of Fraud Detection Model in Real-Life Applications

Integrate our model through an API

- Allow real-time analysis and classification of job postings
- Ensures that fraudulent listings are identified and addressed before they reach job seekers

Periodically review existing listings on platforms

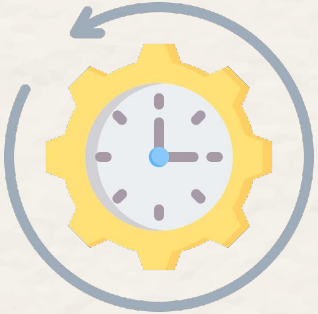
- Safeguarding the job market against scams

Feedback loop

- Continuously improving its adaptability to new fraud patterns.

Limitations

The EMSCAD dataset, while extensive, has a significant class imbalance which can bias the system.



Its temporal scope, focusing on listings from 2012 to 2014, may limit its effectiveness against newer fraudulent strategies

Its limited geographic representation could restrict its applicability globally.

Future Directions

Enrich the dataset with more recent and geographically diverse data

To explore deep learning models for improved accuracy, especially in complex text data analysis

A collaborative effort across job platforms for a unified fraud detection system

Thank You!