Is this Job Real or Fake?

Done By: Joey Kang
DSI23
Capstone Project

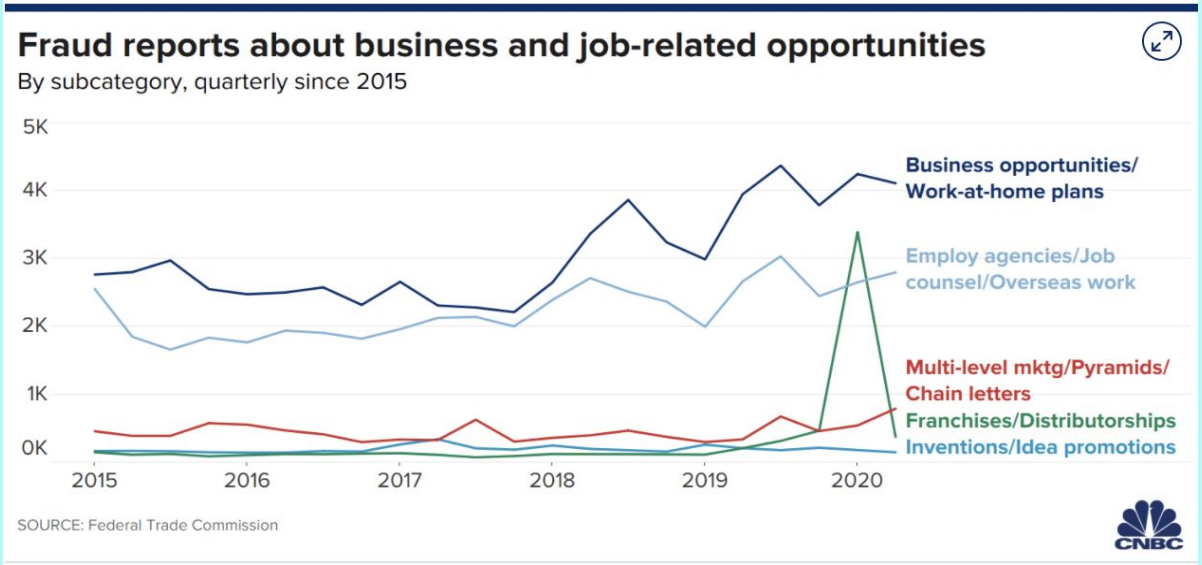
TABLE OF CONTENTS

01
BACKGROUND

BACKGROUND

- COVID-19 has caused unemployment rate to increase due to the economic slowdown
- Rate of employment fraud has **risen by 30%** during the ongoing pandemic period
- **More than 80%** of job seekers report being on guard or very concerned about job scams
- Easy to publish job ads on job portals or on messaging apps

BACKGROUND


Fraud reports about business and job-related opportunities
By subcategory, quarterly since 2015
5K
4K
3K
2K
1K
0K
2015
2016
2017
2018
2019
2020
Business opportunities/ Work-at-home plans
Employ agencies/Job counsel/Overseas work
Multi-level mktg/Pyramids/ Chain letters
Franchises/Distributorships
Inventions/Idea promotions
SOURCE: Federal Trade Commission
CNBC

PROBLEM STATEMENT

Given the increasing number of job scams, we aim to train a classifier to predict whether jobs are real or fake to prevent job-seekers from falling prey to job scams. The classifier will be incorporated to job portals such that if a job listing is predicted to be fraudulent or fake, the listing will not be published on the portal. A successful model would be one with a high ROC AUC score (>0.9) and high recall score.

02 DATA PROCESSING

Exploratory Data Analysis,

Text Pre-processing

DATA SET

- 17,880 rows and 18 columns
  - Include both structured and unstructured data
- Data compiled by the University of the Aegean, Laboratory of Information & Communication Systems Security
- Target Variable: Fraudulent (1 or 0)

DATA PROCESSING

Data Imputation

Backfill categorical columns (i.e. required_experience, required_education, employment_type)

Grouping Categories

Create more generic groups for industry to reduce the number of features

Text Processing

Combined title, description and requirements into a single text column

TEXT PROCESSING

Remove non-english text

Remove links, non-alphabetic characters

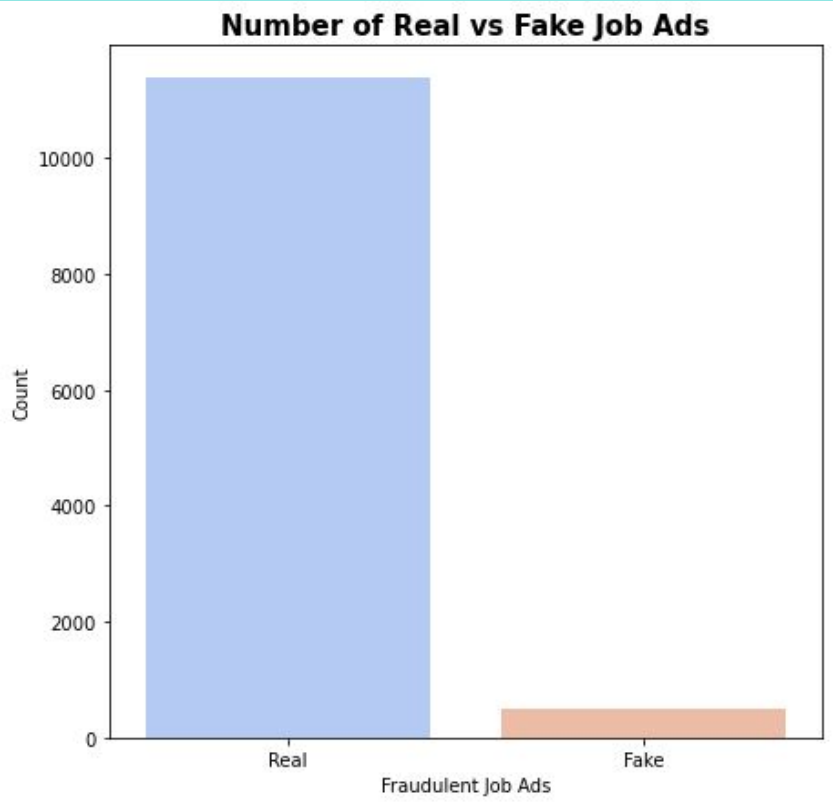
Tokenize and remove stopwords

Lemmatization

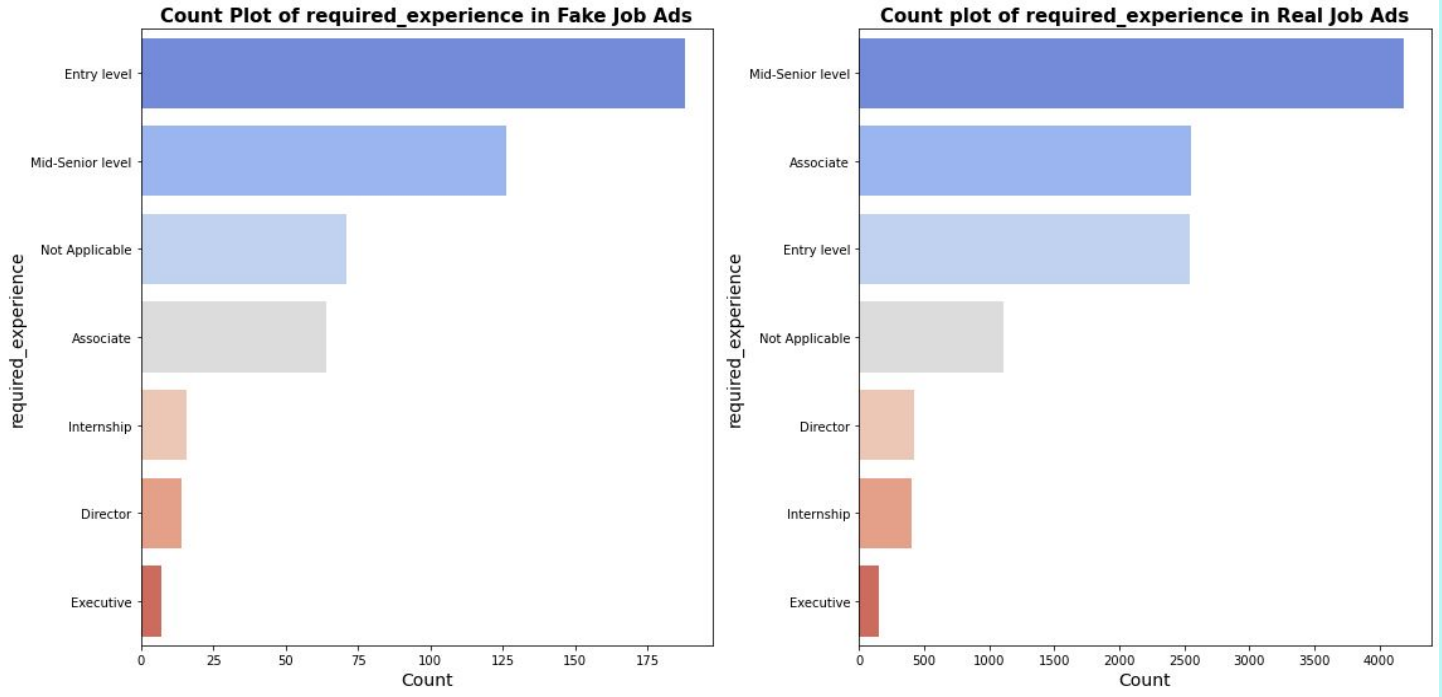
Imbalanced Classification Problem

Majority of the jobs were real (i.e. fraudulent = 0) and only very few were fake (i.e. fraudulent = 1)

May be difficult for the model to predict fake job ads

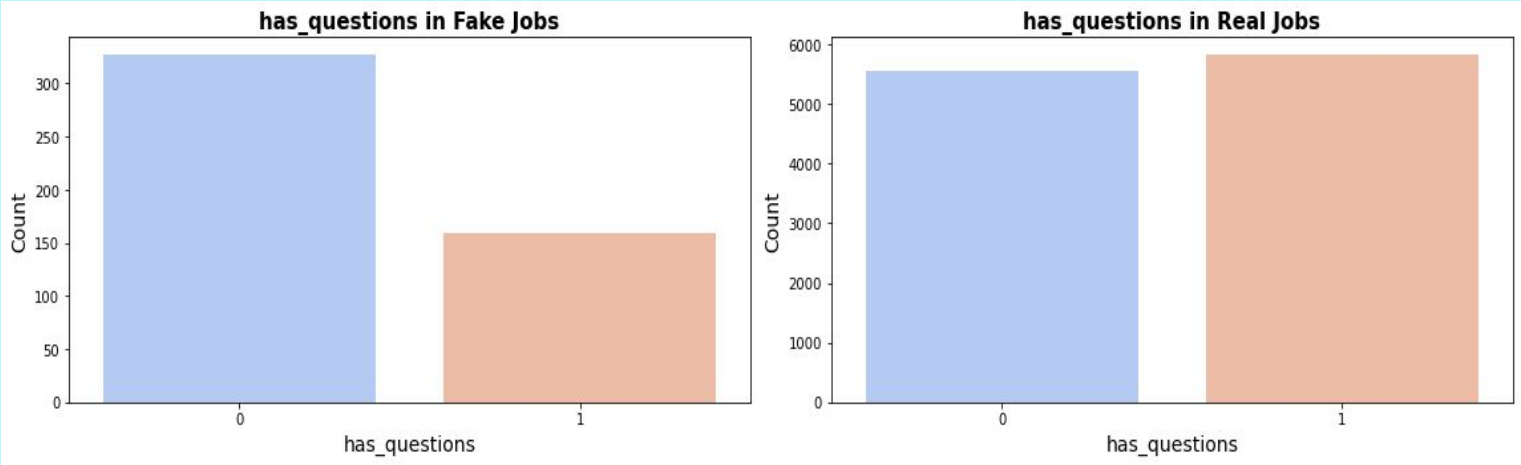


Required Experience



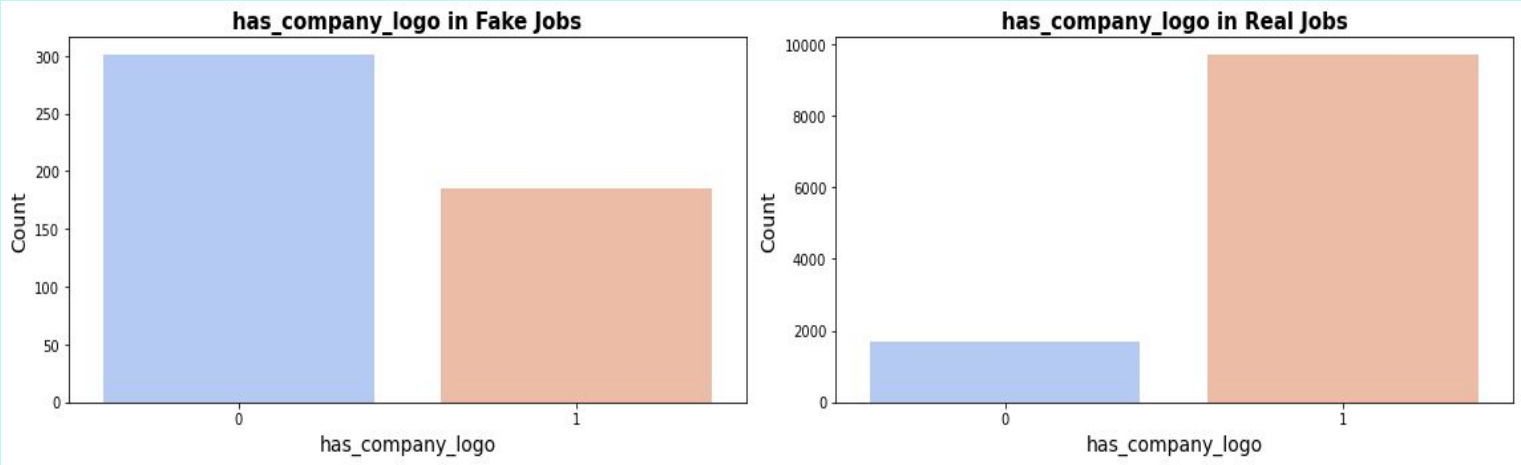
Interview Questions

- Proportion of fake jobs which did not require interview is much higher than those which required interview ⟶ simpler hiring process for fake jobs

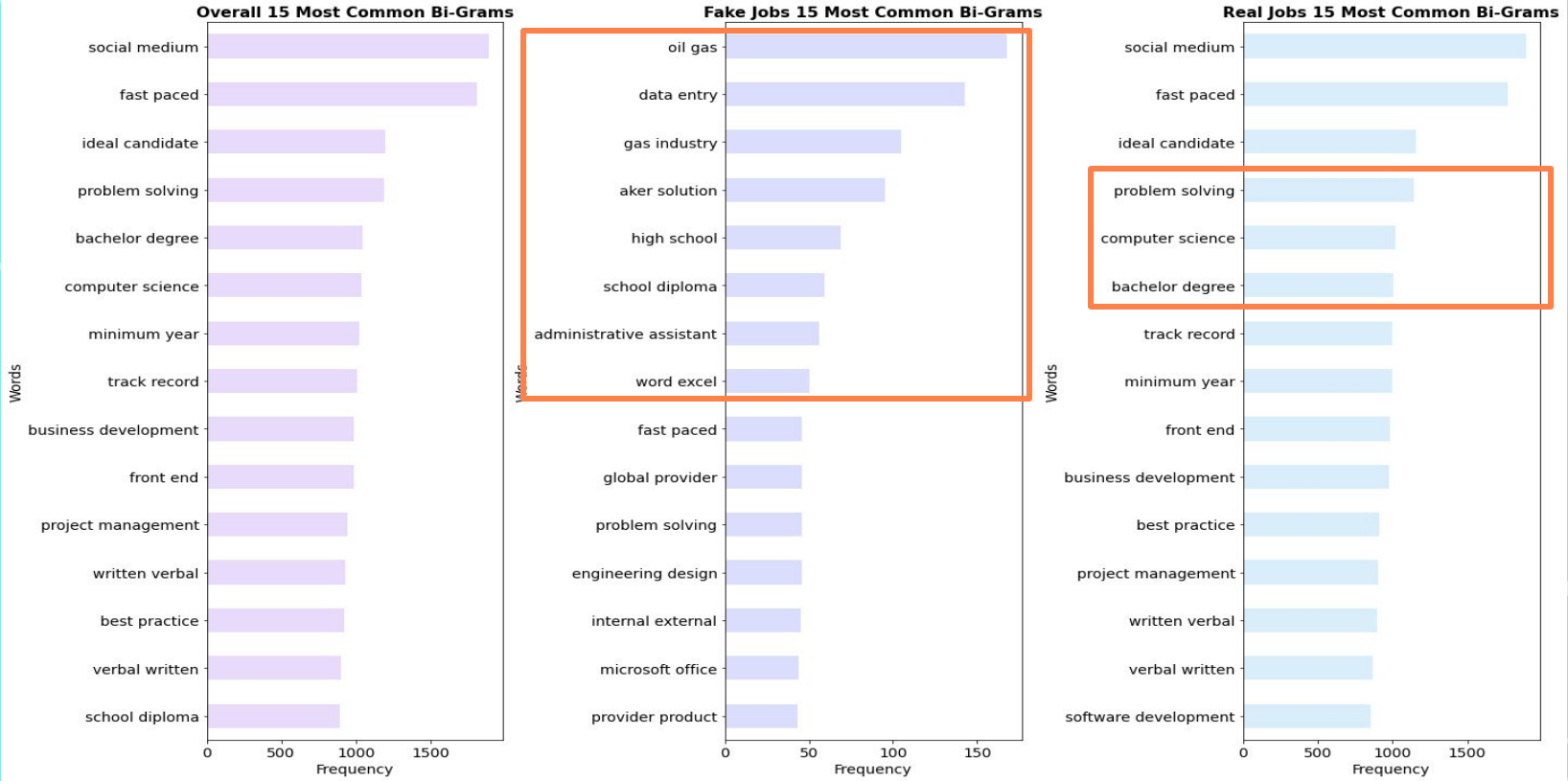


Company Logo

- Most real jobs have a company logo while most fake jobs do not
- Having a company logo increases the credibility of the job



Common Words



03 DATA MODELLING

Model Evaluation

Topic Modelling

MODELLING

Vectorize Text Variables

Dummify Categorical Variables

Include Numerical Variables

Column Transformer → SMOTE → Fit to Model

LOGISTIC REGRESSION

Count Vectorizer

Train Acc: 1.000
Test Acc: 0.966
ROC AUC: **0.925**
Recall: **0.664**

TFIDF Vectorizer

Train Acc: 0.961
Test Acc: 0.948
ROC AUC: **0.946**
Recall: **0.801**

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

EXTREME GRADIENT BOOSTING

Count Vectorizer

Train Acc: 1.000
Test Acc: 0.974
ROC AUC: **0.917**
Recall: **0.582**

TFIDF Vectorizer

Train Acc: 1.000
Test Acc: 0.978
ROC AUC: **0.945**
Recall: **0.610**

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

RANDOM FOREST

Count Vectorizer

Train Acc: 1.000
Test Acc: 0.975
ROC AUC: **0.951**
Recall: **0.459**

TFIDF Vectorizer

Train Acc: 1.000
Test Acc: 0.977
ROC AUC: **0.950**
Recall: **0.466**

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

MULTINOMIAL NAIVE BAYES

Count Vectorizer

Train Acc: 0.962
Test Acc: 0.953
ROC AUC: **0.923**
Recall: **0.664**

TFIDF Vectorizer

Train Acc: 0.922
Test Acc: 0.922
ROC AUC: **0.919**
Recall: **0.712**

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

K-NEAREST NEIGHBORS

Count Vectorizer

Train Acc: 0.927
Test Acc: 0.892
ROC AUC: **0.877**
Recall: **0.760**

TFIDF Vectorizer

Train Acc: 0.928
Test Acc: 0.896
ROC AUC: **0.862**
Recall: **0.760**

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

MODEL RESULTS

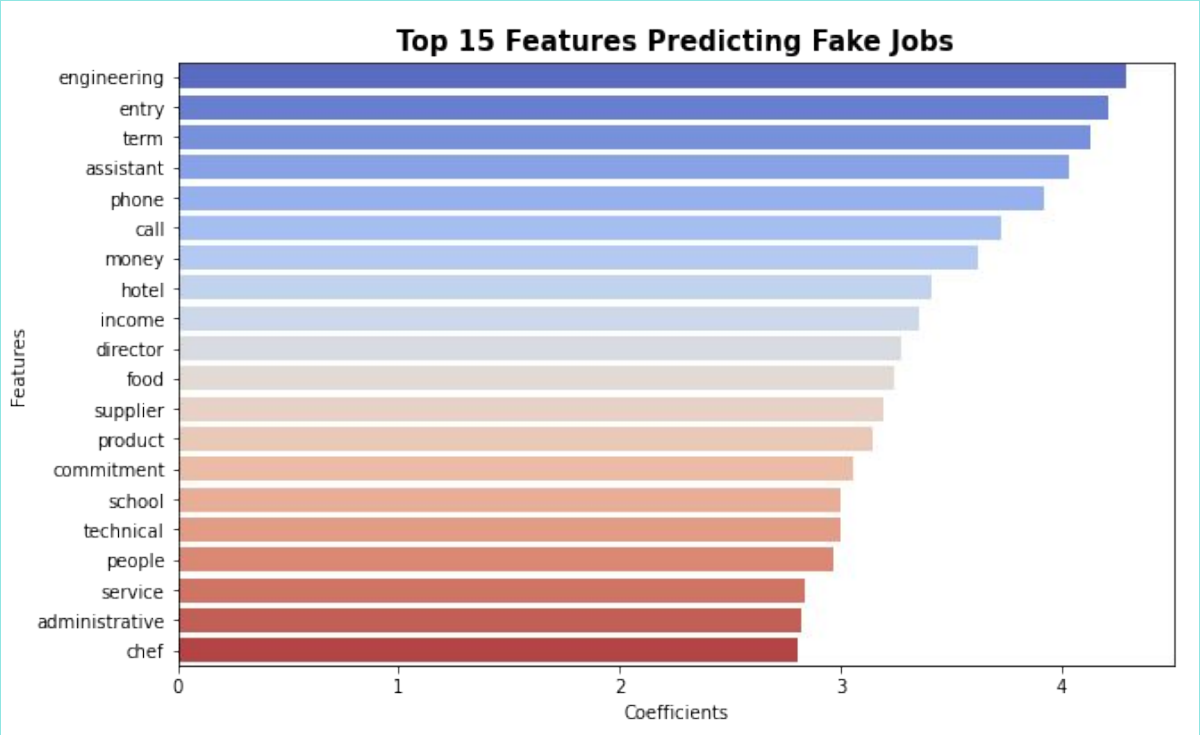
Model	ROC AUC Score	Recall Score
Logistic Regression (tvec)	**0.946**	**0.801**
Extreme Gradient Boost (tvec)	0.945	0.610
Random Forest (cvec)	0.951	0.459
Random Forest (tvec)	0.950	0.466
Multinomial Naive Bayes (cvec)	0.923	0.664
K-Nearest Neighbors (cvec)	0.877	0.760

[1] The closer the ROC AUC is to 1, the better the model is at distinguishing between the two classes
[2] Recall measures the true positive rate: TP/TP + FN

FEATURE IMPORTANCE

Fake Jobs commonly involve:

- Engineering
- Data entry
- Administrative assistant



TOPIC MODELLING

Fake Job ads are clustered around 4 topics, with majority in Topic 0 which is centred around engineering/business management

	Topic	Count	Name
0	0	403	0_system_management_engineering_business
1	1	35	1_position_job_suppliescomputer_distractionsmust
2	2	35	2_care_nursing_patient_rn
3	3	13	3_glass_optical_optician_lens

TOPIC MODELLING

Top 3 topics of real job ads:

- Design
- Quality Assurance
- Java

	Topic	Count	Name
0	-1	3641	-1_digital_want_startup_industry
1	0	221	0_designer_visual_creative_photoshop
2	1	215	1_testing_qa_assurance_automation
3	2	206	2_java_xml_oracle_framework
4	3	195	3_hr_recruitment_recruiting_recruiter
5	4	183	4_manufacturing_maintenance_electrical_repair
6	5	179	5_admin_funding_apprenticeship_na
7	6	168	6_accounting_financial_accountant_finance
8	7	166	7_sql_database_oracle_server
9	8	163	8_ui_designer_visual_designing
10	9	156	9_office_assistant_executive_calendar
11	10	138	10_campaign_communication_advertising_brand

04

CONCLUSION

Conclusion

Recommendations

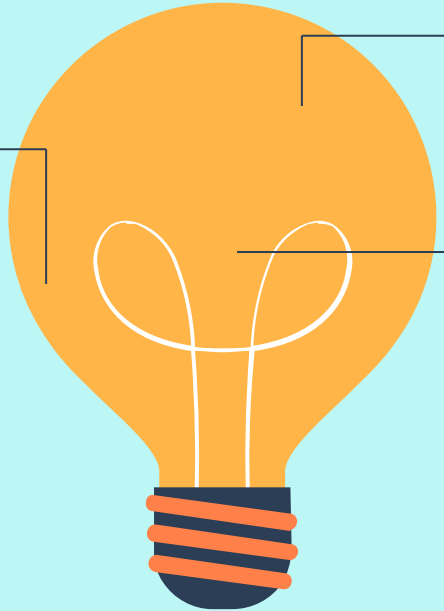
MODEL IMPROVEMENTS

Model correctly predicts the fraudulent jobs 80% of the time and prevents them from being published on job portals → less job-seekers falling prey to job scams

Additional/updated data on fraudulent jobs or additional features such as platform the ad was posted on

Further fine-tune hyperparameters, understand context before removing words or try other oversampling techniques

RECOMMENDATIONS



Conduct background checks on the company

If the job posting does not come with a company logo/profile, do some research on the company (ask around/check online for reviews)

Be wary of 'too good to be true' jobs

If the job is high paying and technical but has low requirements, be wary

Look for jobs with more specialised skills

Fraudulent jobs generally require basic skills, to attract job-seekers who can easily fulfil this criteria

THANKS!

Happy Graduation!