Fraudulent Job Detection

DBA5106 Final Group Project

Prepared by Team 7:

ARIAN MADADI	(A0231939X)
BHARGAV SAGIRAJU	(A0262798J)
LEE CHENG SHU	(A0262750H)
MANAN LOHIA	(A0262838U)
NGUYEN MINH HIEU	(A0262807B)

Table of Contents

Introduction

With the recent global Great Resignation, a record number of people have redirected their career since the start of the pandemic. With the help of online advertisements and job offers, job seekers were able to search for new career opportunities through virtual job hunting. Although the recruitment process is enhanced with the use of online career portals and social media platforms, it has also resulted in uncertainty among job applicants due to the rise of fraud job postings.

Fake online job advertisements are often reflected as job scams that result in stealing of personal and professional information. According to a recent survey by ActionFraud from UK, more than 67% of job seekers are at great risk because of the increased propensity to search for jobs online, without knowing about job scams (Alharby&Alghamdi, 2019). In UK, the report showed almost 300% increase of job scams over 2 years' time (Habiba&Islam&Tasnim, 2021).

In the recent years, scams have continued to be on the rise and acted as the main driver of Singapore's crime in the first half of year 2022. A total of $227.8 million was lost to top 10 scam types, a 60% increase as compared to the same period last year (SPF, 2022). Among different types of scams, job scams topped the list with over 3,573 cases reported and more than $58.5 million lost, according to mid-year crime statistics released by Singapore Police Force (SPF, 2022).

Project Scope

Problem Statement

There are various types of job scams, which involve stealing of personal information such as bank details, national ID, date of birth etc., or advance fee scams where frauds request for payment upfront through reasons such as administrative charges, management cost etc. There were also cases where job seekers, especially students, unknowingly became money mules through paying money into their accounts and transfer it back to scammer's account. Job seekers are often tricked by fake company websites, cloned official looking documents, and were approached by scammers through email or social media such as LinkedIn, disguised as recruitment agencies or headhunters (Habiba&Islam&Tasnim, 2021).

Given the huge number of online job postings, job seeking experience will be greatly improved if job posts could be filtered, with only legitimate job postings remaining on job searching platforms. Fake job posts could result in a huge waste of time and resources, and unintentional personal information disclosure for job seekers. Job scam also has a negative impact on recruitment agencies as it damages business reputations and create negative image for agencies. Randstad was targeted with job scams when individuals and/or organizations were impersonating Randstad employees and scam people with fake job offers (How to avoid and report job scams, 2022). Therefore, it is important to know what to look out for in a job posting and how to identify fake job postings that might be a scam to steal personal information.

Project Objective

This study aims to develop and evaluate models that can identify if an online job posting is real or a scam. Through this study, our group has identified the following questions to address the problem statement and our objective:

Research Question 1: Which key features in the job description are effective to detect fraudulent job postings?

Research Question 2: Which model is the most effective in predicting if a job posting is real of fake?

The final proposed model will be able to take in job posting data with relevant features and predict whether it is real or fake job postings.

Methodology

This study follows the general workflow of the business analytics cycle, with business understanding defined as a start and constantly checked throughout the process to ensure that identified business problem statement is addressed, as illustrated in Figure 1.

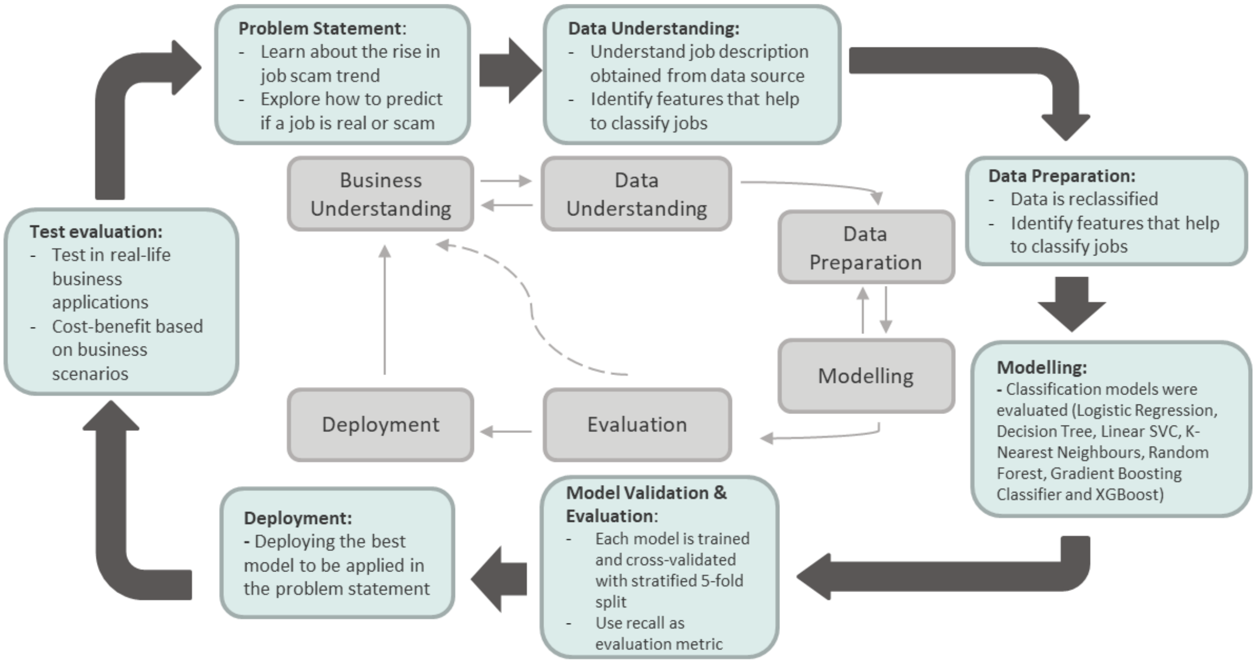

Problem Statement:
- Learn about the rise in job scam trend
- Explore how to predict if a job is real or scam
Data Understanding:
- Understand job description obtained from data source
- Identify features that help to classify jobs
Data Preparation:
- Data is reclassified
- Identify features that help to classify jobs
Modelling:
- Classification models were evaluated (Logistic Regression, Decision Tree, Linear SVC, K-Nearest Neighbours, Random Forest, Gradient Boosting Classifier and XGBoost)
Model Validation & Evaluation:
- Each model is trained and cross-validated with stratified 5-fold split
- Use recall as evaluation metric
Deployment:
- Deploying the best model to be applied in the problem statement
Test evaluation:
- Test in real-life business applications
- Cost-benefit based on business scenarios
Business Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Figure 1. *General workflow of the business analytics cycle in this study*

Data Preparation and EDA

Data

Data is extracted from Kaggle (Chauhan, 2022), containing roughly 18,000 job listings, out of which about 800 are fake. The data consists of a combination of integer, binary and textual data types. This study aims to create classification models, which identify if the job listings that are fraudulent. Data variables are described in Table 1 below.

***Table 1**: Data Variables Description*

Variable	Data Type	Description
job_id	Int	Unique Job ID
title	Text	The title of the job ad entry
location	Text	Geographical location of the job ad
department	Text	Corporate department (e.g., sales)
salary_range	Text	Indicative salary range (e.g., $50,000-$60,000)
company_profile	Text	A brief company description
description	Text	The details description of the job ad
requirements	Text	Enlisted requirements for the job opening
benefits	Text	Enlisted offered benefits by the employer
telecommuting	Boolean	True for telecommuting positions
Has_*company*_logo	Boolean	True if company logo is present
has_questions	Boolean	True if screening questions are present
employment_type	Text	Full-type, Part-time, Contract, etc
required_experience	Text	Executive, Entry level, Intern, etc
required_education	Text	Doctorate, Master's Degree, Bachelor, etc
industry	Text	Automotive, IT, Health care, Real estate, etc
function	Text	Consulting, Engineering, Research, Sales etc
fraudulent	Boolean	target - Classification attribute

Data Preparation and Feature Engineering

The dataset is explored to identify peculiarities and prepare the dataset for EDA and modelling. An initial look at the data shows us that most of the values (roughly 80 %) for the *salary_range* are missing. Hence, we choose to drop the column. Additionally, *job_id* is also dropped as it is an ID column. From the *location* field, we extract the country for which the job is being offered and drop the State and the City. Nonetheless, before dropping the city, we identify listings which have multiple locations and create a flag variable for the same. For the other non-categorical text columns, namely, *department*, *company_profile*, *description*, *requirements,* and *benefits*, we engineer Boolean flag variables to indicate if a value for each row exists or is missing.

As there are several industries for which jobs are posted, we have bucketed all industries not in the top 12 by count into an 'Other' category, based on the total counts of each industry. This gives us a total of 13 categories under the industry column, as shown in Figure 2

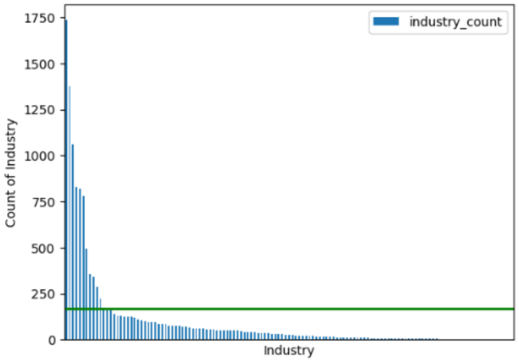


Figure 2. *Industries by count. The Green line shows the cutoff for the 'Other' category*

Finally, we use One-Hot-Encoding to convert all our remaining categorical text features into Boolean indicator variables so that our models can work with them. As we intend to work with a variety of models, some of which are capable of implicitly handling missing values, we do not drop any rows with missing values and neither perform imputation of any sort.

The dataset is imbalanced, with more than 95% of the jobs being real and only 4.8% of fraudulent jobs. Figure 3 shows the clear difference in count between real and fraudulent jobs.

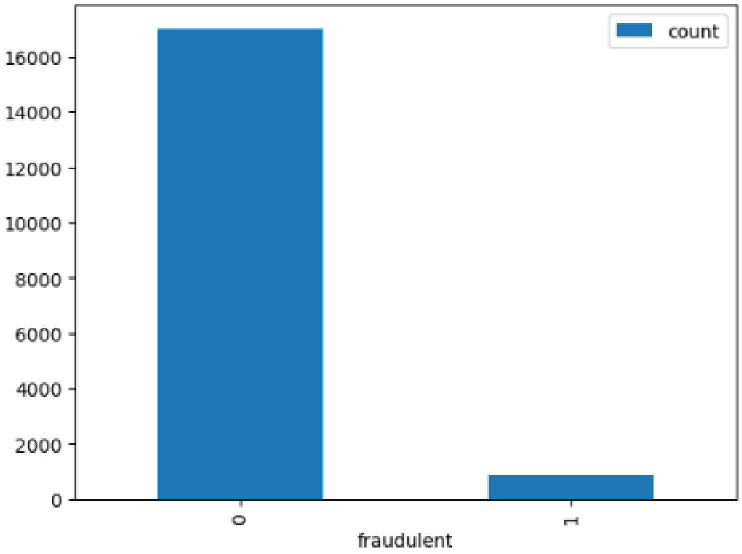


Figure 3. *Count of fraudulent (1) and non-fraudulent (0) job listings*

Exploratory Data Analysis (EDA)

Having cleaned the data, we moved ahead to looking at the data to find any peculiarities or insights that could be of interest before we start building a model. The first thing to note is that our dataset consists only of categorical columns and no numerical columns. This suggests that tree-based models could be very effecting in helping us identify fraudulent jobs.

We also notice a disproportionate number of fraudulent cases wherein the job listing does not have a company profile or company logo. The same is evident in the plots in Figure 4.

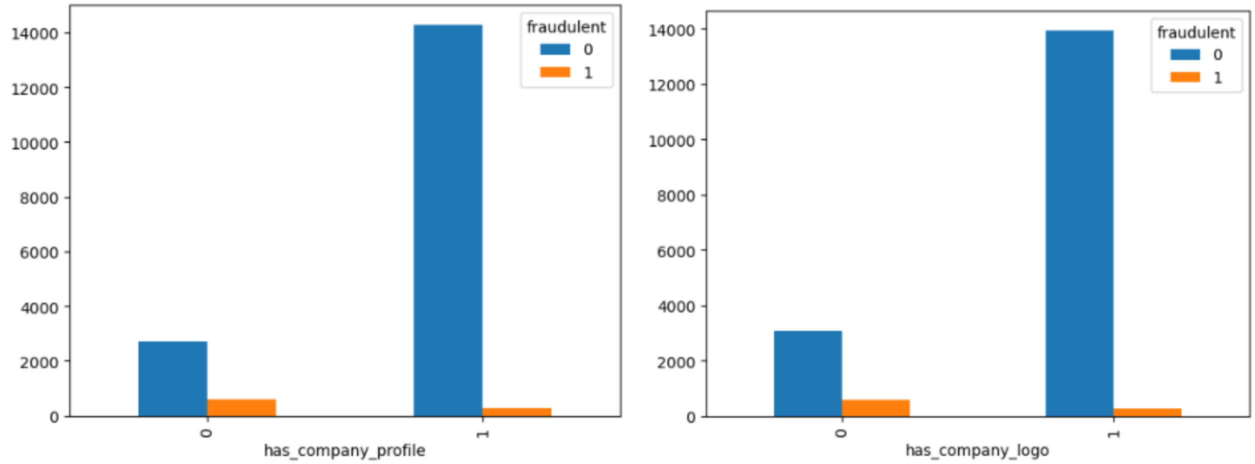
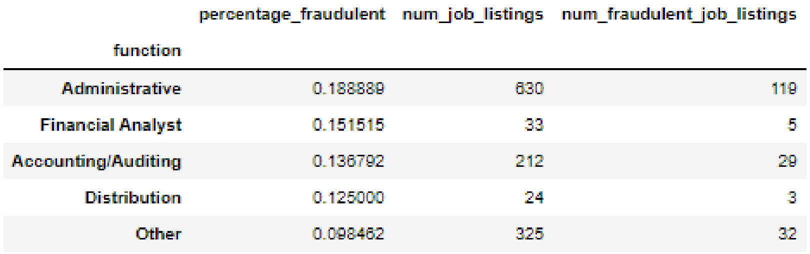


Figure 4. *Count of fraudulent (1) and non-fraudulent (0) job listings by has_company_profile (left) and has_company_logo (right)*

Listings for jobs at the executive or entry level, or part time jobs appear to have the highest proportion of fraudulent jobs in their respective categories. However, as we only observe a small sample of all the job listings, this observation might be skewed. Along the same lines, job functions with the highest amount of fraudulent job listings are Administrative, Financial Analyst, Accounting/Auditing and Distribution.



function	percentage_fraudulent	num_job_listings	num_fraudulent_job_listings
Administrative	0.188889	630	119
Financial Analyst	0.151515	33	5
Accounting/Auditing	0.136792	212	29
Distribution	0.125000	24	3
Other	0.098462	325	32

Figure 5. *Top 5 job functions by percentage of fraudulent listings*

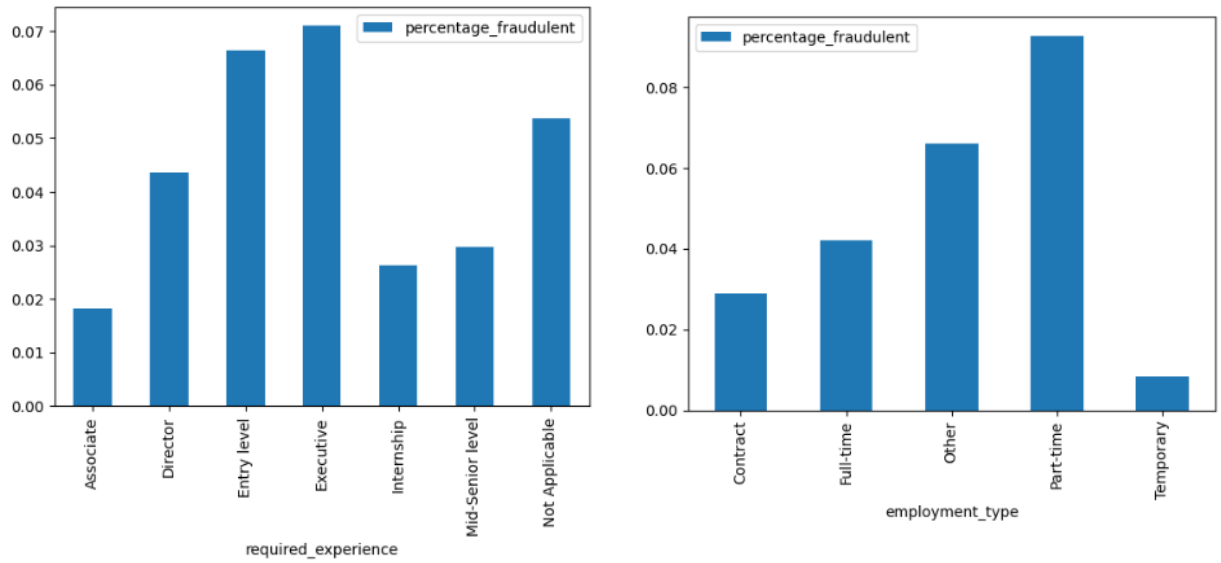


Figure 6. *Percentage of fake jobs by required experience (left) and employment type (right)*

Model Development

For the model, we decided to use the following classification models: Logistic Regression, Decision Tree, LinearSVC, K-Nearest Neighbors, Random Forest, Gradient Boosting Classifier and XGBoost. We fit each of these models over our dataset containing 299 features in total. Since the classes are imbalanced, we add class weights to allow the models to balance the predictions. We then cross-validate each of these models using a stratified 5-fold split and score them on average Recall. We find that Gradient Boosting Classifier has the best Recall score of 0.88 and on further investigation has a low fluctuation in the Recall across each fold. This is interesting because tree-based models are very effective when it comes to working with several features; however, identifying fraudulent jobs in this case is more effective with a significant number of weak learners.

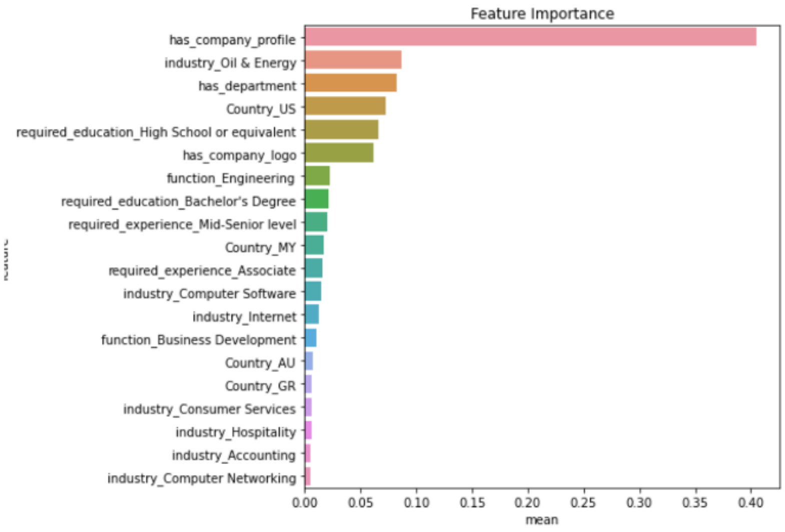


Figure 7. *Top 20 important features*

Model hyperparameter tuning and performance evaluation

We proceed to perform grid search to tune the hyperparameters which results in a model with a Recall score of 0.84 for the fraudulent class, with the best parameters being 50 estimators each with a learning rate of 0.1 and a max depth of 3.

Table 2.1: *Model Performance and Comparison*

Model	Average Recall
Gradient Boosting Classifier	0.88
Decision Tree	0.77
LinearSVC	0.71
Logistic Regression	0.71
Random Forest	0.70
XGBoost	0.59
KNN	0.59

Table 2.2: *Gradient Boosting Classifier performance across folds*

fold	fit_time	score_time	test_recall
1	1.78	0.01	0.89
2	1.78	0.01	0.85
3	1.78	0.01	0.88
4	1.78	0.01	0.88
5	1.81	0.01	0.90

Model Evaluation

For our model evaluation, we choose to use Recall as our evaluation metric.

$$Recall = TPR = \frac{TP}{P}$$

The model intends to identify as many True Positives as possible, even if the algorithm will have many False Positives as a result. The intuition behind is that if a job posting is flagged as fake, the company listing the job would simply need to provide additional information/verification to the job listing service. For a real job (False Positive), this is a simple task and would not negatively impact them too much. Nevertheless, for a fake job (True Positive), this would be difficult to provide, and ideally result in the job listing simply being withdrawn or flagged as fake on the website.

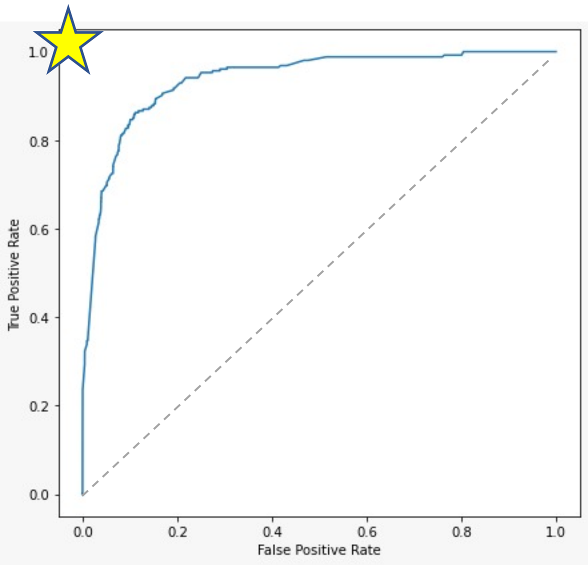


Figure 8. *ROC curve*

Based on the ROC curve plotted in Figure 8 above, our model seems to be performing much better as compared to the “no skill” classifier, represented with dotted line. However, it was mentioned that ROC curve is not recommended for highly imbalanced data, and our data source was identified as imbalanced. Therefore, we will further explore our model through confusion matrix + cost-benefit matrix, to obtain the expected value, in order to get a feel of how our model is performing, with business assumptions.

Confusion Matrix

Based on the selected model (Gradient Boosting Classifier), we have constructed the confusion matrix with numbers and probability as listed in table 3.1 and 3.2.

Table 3.1: *Confusion Matrix (numbers) for this study*

	Predicted Real	**Predicted Fake**
Actual Real	4511	599
Actual Fake	35	219

Table 3.2: *Confusion Matrix (probability) for this study*

	Predicted Real	**Predicted Fake**
Actual Real	0.84	0.11
Actual Fake	0.01	0.04

Cost-Benefit Analysis

As mentioned in the problem statement, other than resulting in a huge waste of time and resources for job seekers, fake job postings lead to certain economic loss by recruitment agencies due to business reputational damage. The most common method to monetize a job board website is through charging employers for job postings, promoting job listings, charging viewers to access certain job details, gaining revenue through advertising space and promoting additional services through subscription methods (Andrushchenko, 2022). For the job board to be profitable, it should fulfill criteria such as decent website traffic and a large database of credible employers (Andrushchenko, 2022). If a job board website is identified with large amount of fake job postings, it will affect the website's credibility, reducing its traffic and premium subscription, which in turn affect its earnings. Employers will then utilize more credible websites for their recruitment, reducing the job board website's revenue. Cost-Benefit Matrices are generated as shown in Table 3.3 and 3.4 below.

Assumptions made in generating the Cost-Benefit Matrix are listed as below:

- Job board websites only gain revenue through 3 methods:
    - Recruitment services (generated from employer based on cost-per-click)
    - Advertisement (generated when user clicks into the job listing)
    - Premium subscriptions
- Earning generated per actual real click through recruitment services is $5
- Earning generated per predicted real click through advertisement is $2
- Earning generated through premium subscription is $5
- Cost of maintaining website per user is $2; there is no other cost involved
- Processing cost for additional verification when real job is listed as fake is $2
- Penalty cost due to actual fake listing identified as real is $5
- Job seekers who clicked on actual real and successfully identified actual fake will subscribe to premium subscription with probability of 0.8
- Job seekers will follow predicted results (i.e., click in when it's predicted as real and does not click when it's predicted as fake)

Table 3.3: *Cost-Benefit Matrix analysis based on assumptions made*

	Predicted Real	**Predicted Fake**
Actual Real	Recruitment service earning + Advertisement earning + Premium Subscription – Maintenance cost	– Maintenance cost – Processing cost
Actual Fake	Advertisement Earning – Maintenance cost – penalty cost	Premium subscription – Maintenance cost

Table 3.4: *Cost-Benefit Matrix for this study*

	Predicted Real	**Predicted Fake**
Actual Real	$5 + $2 + $5*0.8 – $2 = $9	-$2 - $2 = -$4
Actual Fake	$2 – $2 - $5 = -$5	$5*0.8 - $2 = $2

Expected Value Framework

Expected value is computed to be $7.15, by combining the model prediction confusion matrix with the cost-benefit matrix. The expected value turns out to be quite significant based on the business assumptions made. However, it might vary according to the various business models applied by the job board websites.

Table 3.5: *Expected Value Framework for this study*

	Predicted Real	**Predicted Fake**
Actual Real	0.84*$9 = 7.56	0.11*(-$4) = -0.44
Actual Fake	0.01*(-$5) = -0.05	0.04*($2) = 0.08

Recommended Use

The intended use case for this model is on job board websites such as Indeed (indeed.com) and JobStreet (jobstreet.com.sg). The websites would implement the algorithm and decide how strict they wish it to be – as increased strictness would increase the number of true positives, but also the number of false positives. The Cost-Benefit Analysis would be left to each individual company (as in the example above). Once the algorithm determines that a job listing is fake, the company (using Indeed as our example), would send an automated message to the job lister, asking for additional verification of the job's validity. A company offering a real job would be able to provide this with little issue, whilst a fake job would not be able to provide it. As a result, the job listing would be taken down, and the poster/poster's company banned from the website.

Conclusion

Based on existing dataset on job posting legitimacy, we have performed data preparation and explanatory analysis to obtain an initial understanding of our dataset. We also develop 7 different classification models (Logistic Regression, Decision Tree, LinearSVC, K-Nearest Neighbors, Random Forest, Gradient Boosting Classifier and XGBoost) to predict fraudulent job listings. Using Recall as our evaluation metric, we conclude that Gradient Boosting Classifier is the best model with a Recall score of 0.88 and generally low Recall fluctuation across folds.

We have made several business assumptions and constructed a Cost-Benefit Matrix, and along with our Confusion Matrix, we build an Expected Value Framework to quantify the business value of our model predictions. As the framework suggests that our predictions have positive commercial value, we recommend our model to job posting organizations. And last but not least, as the data is highly imbalanced, we propose to gather more data for future analysis.

References

Alharby&Alghamdi. (2019). An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, 10. 155-176. doi:10.4236/jis.2019.103009

Andrushchenko, S. (2022). *Top 5 Ways to Monetize a Job Board Website*. Retrieved from https://hivepress.io/blog/top-5-ways-to-monetize-a-job-board-website/#models

Chauhan, A. (2022). *Real OR Fake Jobs*. Retrieved from https://www.kaggle.com/datasets/whenamancodes/real-or-fake-jobs

Habiba&Islam&Tasnim. (2021). A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques. *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST).* doi:10.1109/ICREST51555.2021.9331230.

How to avoid and report job scams. (2022). Retrieved from https://www.randstad.com.sg/job-seekers/job-scams/

SPF. (2022). *Mid-Year Crime Statistics 2022.*