Group 4 Final Report
Beware of Job Scams! Fake Job Prediction

Group 4; Lam Fu Yuan, Kevin (A0094651B); Lee Hui Gek Judy (A0218942Y); Lim Wei Ren (A0218875N); Mansi Agarwal (A0218968J); Zhang Jie (A0231865B)

1. Introduction

1.1 Background

As Singapore advances into a Smart Digital Nation, digitization of services has enhanced our convenience, but also act as a double-edged sword as they expose users to the risk of scams. During the Covid-19 pandemic, adoption of contactless digital payments increased tremendously and created a large pool of targets for scammers.

While many people were displaced from their jobs due to the pandemic, the incidences of reported job scams skyrocketed due to the lure of jobs advertised to be convenient and high paying. An instance of such scams misled victims to websites that enticed them to subscribe packages and to transfer money to unknown bank accounts to be paid the commission (Lim & Sun, 2022). Alternatively, scammers use fake job listings to trick victims into disclosing personal information, allowing them access to victims' online credit cards or bank accounts (Rafter, 2021). Straits Times reported that in the first six months of 2021, there were 658 cases of job scams, a 16-fold increase from 40 in the same period of 2020. During these six months, victims lost ~S$6.5 million, which is a considerable increase from $60,000 during the same period in 2020 (Sun & Lim, 2022).

Large international companies are not spared from recruitment fraud. Since early 2020, video game giant Riot Games has been trying to deal with scammers who lured eager professionals into handling sensitive data by dangling fraudulent job postings. Some recruiting websites even allowed potentially false jobs to be posted on an official company page, appearing next to legitimate listings. The US Federal Bureau of Investigation estimate these scams to cost victims an average of US$3,000 and often cause long term harm by negatively impacting victim's credit scores (Janofsky, 2022). The scale and impact of this problem has given recruitment giants such as Linkedin impetus to act. Between Jan to Jun 2021, LinkedIn removed 66.3 million spam or scam content, of which 99.6% of these were stopped by automated defenses (Linkedin, 2022).

1.2 Problem Statement

Back in Singapore, Government Agencies have developed "ScamShield" that helps to block scam calls and filter scam messages (Scamshield, n.d.). However, there is no tool widely available to the job seeking population for discerning fake job listings.

We believe machine learning can be effectively applied to tackle this problem by training a model for fake job classification. For the purpose of this project, we will be focusing on job scams from job portal sites where it is more difficult to determine authenticity compared to questionable job offers received via messaging platforms such as Whatsapp.

1.3 Business Value

Coming from different perspectives, we have summarized the benefits which the different audiences can reaped in Table 1. With a proven model design, we can extend the usage to detect job portals / sites that have a high percentage of potential fake jobs. This would be an attractive tool for search engines.

Table 1. Business Value for Different Audiences

For Job Seekers	***For Job Portals***	***For Job Posters***
Benefits		
• Reduce time wasted • Minimize risk of phishing • Increased confidence	• Improved accuracy in detecting fraudulent postings • Increased platform trust • Increased user growth / stickiness • Attractiveness to partners	• Reduced false positives from existing flagging mechanisms
Willingness to Pay		
Low	High	Medium

2. Research Questions

We aim to answer the following three research questions:

RQ1: Which features are most associated with fraudulence?

RQ2: What are the topics present in the company profile, job description, job requirements and job benefits?

RQ3: Which machine learning model is most suitable to predict fraudulence

Table 2. Description of Variables in the Employment Scam Aegean Dataset (N = 17,880).

No.	Name	Type	Description
1	Job ID	Nominal	Serial number (e.g., 1).
2	Job Title	String	Title (e.g., Infrastructure Engineer).
3	Job Location	Nominal	Geographical location (e.g., US, CA, California).
4	Job Department	Nominal	Corporate department (e.g., Accounting).
5	Job Salary	Ordinal	Indicative salary range (e.g., $7000-$9000).
6	Company Profile	String	Company description (e.g., "Our mission to clients is …").
7	Job Description	String	Job description (e.g., "Drive the sales effort …").
8	Job Requirements	String	Job requirements (e.g., "Proven leadership experience …").
9	Job Benefits	String	Job benefits (e.g., "Fun, supportive team …").
10	Telecommuting	Nominal	Telecommuting position (1 = Yes; 0 = No).
11	Company Logo	Nominal	Presence of company logo (1 = Yes; 0 = No).
12	Questions	Nominal	Presence of screening questions (1 = Yes; 0 = No).
13	Employment Type	Nominal	Employment type (e.g., Full-Time, Part-Time).
14	Required Experience	Ordinal	Required experience (e.g., Entry-Level, Mid-Senior Level).
15	Required Education	Ordinal	Required education (e.g., Bachelor's Degree, Master's Degree).
16	Job Industry	Nominal	Industry (e.g., Banking, Design).
17	Job Function	Nominal	Function (e.g., Administrative, Advertising).
18	Fraudulent	Nominal	Fraudulence (1 = Fraudulent; 0 = Not Fraudulent).

Notes. A nominal variable is a variable whose values are treated as categories without a hierarchical ordering; an ordinal variable is a variable whose values are treated as categories with a hierarchical ordering; and a string is a variable whose values are treated as text.

3. Data

3.1 Data Source

The Employment Scam Aegean Dataset (EMSCAD) was downloaded from the Laboratory of Information and Communication Systems Security, Department of Information and Communication Systems Engineering, at the University of Aegean[1]. The EMSCAD contains 17,780 job advertisements published on a recruitment software between 2012 and 2014; each job advertisement was classified as either fraudulent or not fraudulent by specialized employees of the recruitment software[2]. Table 2 describes both the features and the target variable (Fraudulent) in the dataset.

3.2 Data Pre-Processing

The data was pre-processed in three steps. In the first step, some nominal variables such as employment type, required experience and required education were bucketed to reduce the granularity. In the second step, all variables were processed to indicate whether a sample is missing a value on each of them (1 = Missing; 0 = Not Missing). This create a new "missingness" feature. In last step, we concatenated strings from department, company profile, description, requirements, benefits, industry and function sections into a single feature. After having concatenated the strings, we applied the following steps in order to pre-process them:

- Convert all alphabets to lowercase
- Remove all punctuations
- Remove all numbers
- Remove all stopwords
- Remove all non-English words
- Remove all rare words
- Lemmatise words

[1] http://emscad.samos.aegean.gr/

[2] https://www.mdpi.com/1999-5903/9/1/6/htm

These steps improved the performance of the LDA model. Only one out of 17,880 samples were removed as a result of the application of these steps. The sample was removed because it contained too few words: with "Office Manager" as its title, there was no other information in the department, company profile, description, requirements, benefits, industry and function sections. Weightage of each topic generated from topic modelling were added into the dataset as new features for machine learning (except for CNN and BERT).

For Convolutional Neural Network (CNN) modeling, all text features were combined and tokenized to words which were converted to numbers via text to word sequence. Word sequencing forms the input features for CNN.

3.3 Exploratory Data Analysis

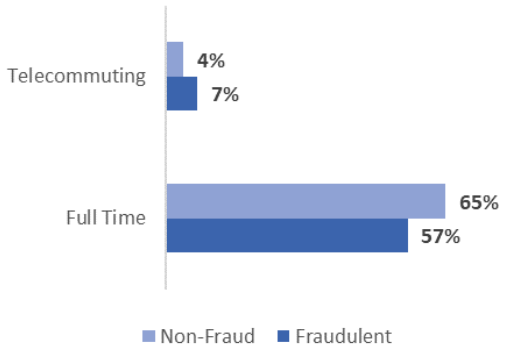
Initial exploratory data analysis was conducted to examine the distribution of the target variable, the associations between values on the feature variables and those on target variable, and the associations between missingness on the feature variables and values on the target variable. Some interesting insights found are:

First, the distribution of the target variable is skewed (Fraudulent: 866 [5%], Non-Fraudulent: 17,014 [95%]). This suggests that we should consider techniques including but not limited to random oversampling or synthetic minority oversampling technique (SMOTE).

Second, the presence of company profiles, company logos and screening questions, whether the job was a telecommuting position and the employment type were associated with fraudulence. Fraudulent job posts mostly lacked company profiles and company logos, and these posts also eradicated the need for screening questions in the survey, portraying an easy-to-apply hiring process to attract more applications.

Hence, we notice fraudulent job posts were more likely to be telecommuting positions (Fraudulent: 64 [7%], Non-Fraudulent: 703 [4%]), and were less likely to be full-time positions (Fraudulent: 490 [57%], Non-Fraudulent: 11130 [65%]).

Figure 1: Comparison of proportions as a % of total fraudulent / non-fraudulent jobs respectively

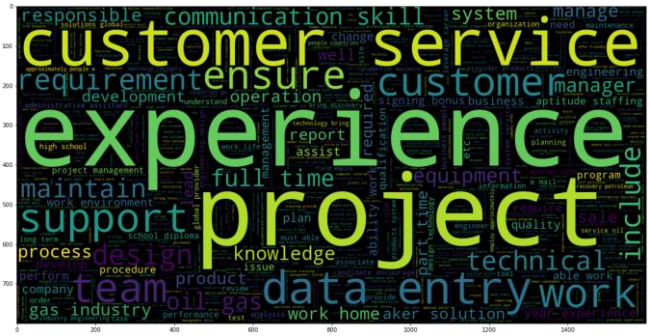


Third, a considerable number of features suffer from a high percentage of missing values (e.g., Job Salary: 84%, Job Department: 65%, Required Education: 45%, Job Benefits: 40%, Required Experiment: 39%). It is interesting that fraudulent job posts were more likely to provide salary information (Fraudulent: 223 [26%], Non-Fraudulent: 2645 [16%]). This is consistent with our domain understanding that most legitimate companies have policies that discourage the presentation of such information in public domains. For example, job salary is confidential and not disclosed to avoid competitors setting a higher benchmark.

Last, given that the dataset contains 5 features of string types (Job Title, Company Profile, Job Description, Job Requirements and Job Benefits), we attempted to tokenize these text strings into words and created word clouds as shown in Figure 2 (Non-Fraudulent) and Figure 3 (Fraudulent). Based on preliminary visual analysis, word such as "full time" appears more frequently in non-fraudulent job listings than fraudulent job listings. Another interesting insight is the word "experience" and "project" have a high count while "customer service" seems to be appear as frequent in both non-fraudulent and fraudulent job listings.

Figure 2: Non-Fraudulent Word Cloud

Figure 3: Fraudulent Word Cloud



4. Topic Modelling

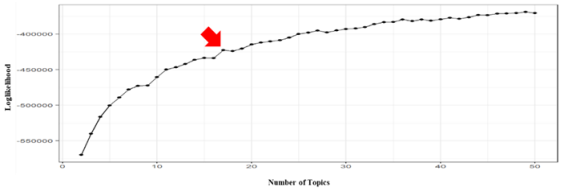
Before employing supervised models, we used Topic Modelling, an unsupervised ML technique which is a

quick way to gather the cluster words and does not require training. Topic models are algorithms that enable the discovery of hidden topical patterns or thematical structures in a large collection of documents. Latent Dirichlet Allocation (LDA) is one such algorithm. LDA seeks to maximise the separation between the means of projected topics and minimise the variance within each projected topic. The topic model allows us to deduce what each set of texts are talking about. A limitation is that topic modelling does not guarantee accurate results.

4.1 Model Selection

The optimal number of topics was determined using the elbow method. Because the elbow method involves a grid search over multiple numbers of topics which has high computational cost, it was conducted on a random subset of 1,000 job advertisements. After having identified the optimal number of topics, the LDA model was then fit on the entire dataset. The results of the grid search suggests that a model with 17 topics would fit the data sufficiently well (Figure 4).

Figure 4. Determining the optimal number of topics



4.2 Model Interpretation

For each topic, we looked at the top 5 most relevant terms to deduce what each set of texts represents. In general, the topics were associated with the sector, the company or the skills of the corresponding job advertisement.

Topics 1, 3, 10, 16 and 17 are associated with the sectors data, web development, healthcare, manufacturing and education respectively. Topics 2, 4, 5, 6, 7, 8, 9, 12 and 14 are associated with the companies Upstream, Spectrum Learning, Tidewater Finance Co., Ixonos, Applied Memetics, ABC Supply Co., Vend, Network Closing Services and Novitex Enterprise Solutions. Topics 11, 13 and 15 are associated with the skills communication, stakeholder co-ordination and teamwork respectively. Table 3 summarises each of the 17 topics based on their most representative observations and words.

Table 3. Summary of the Topics

Topic	Feature Impt	Top 5 Job IDs	Top 5 Words Associated	Topic Description
1	4	2722, 10351, 10050, 16950, 377	system, data, support, inform, technic	Jobs related to system and data
2	16	1304, 8737, 12421, 320, 12098	will, work, new, learn, within	Jobs related to “Upstream” company
3	7	905, 2228, 3086, 1565, 3866	develop, web, test, use, end	Jobs related to web development
4	13	10775, 414, 7170, 14760, 9396	recruit, candid, will, career, avail	Jobs related to “Spectrum Learning” company
5	1	2973, 7188, 14937, 10084, 2170	benefit, employ, employee, posit, paid	Jobs related to “Tidewater Finance Co.” company
6	14	9551, 14218, 7279, 11402, 9263	design, product, project, work, team	Jobs related to “Ixonos” company OR jobs related to UX design
7	5	3950, 3826, 2977, 3078, 3940	market, brand, media, digit, social	Jobs related to “Applied Memetics” company
8	15	17181, 13224, 289, 12592, 669	sales, custom, product, will, return	Jobs related to “ABC Supply Co.” company
9	12	10386,7500,10443,11704,10342	work, great, make, look, can	Jobs related to “Vend” Company OR a Company profile that has lot of office benefits
10	2	6805,173,12730,17230,1224	care, home, train, health, assist	Jobs related to Healthcare
11	11	17331,14128,1280,4823,4663	skill, excel, strong, work, must	Jobs that require strong communication skills
12	17	17060,16988,16889,17031,17009	client, account, profession, close, success	Jobs related to “Network Closing Services” company that deals with settlement of property transactions
13	8	2616,4308,10078,16776,9822	plan, organ, develop, maintain, report	Jobs that require planning and coordination with stakeholders
14	3	9467,11145,1920,12505,3018	custom, process, perform, document, product	Customer Service Role with “Novitex Enterprise Solutions” company
15	10	7957,11676,7445,4375,1498	team, grow, build, fast, lead	Team based roles related to growth in partnerships / clients
16	9	13861,2367,9873,10283,17135	job, will, high, time, posit	Full time positions in Manufacturing
17	6	1232,1983,4888,6056,6976	amp, help, job, get, month	Jobs related to overseas educator

5. Classification Model Selection and Evaluation

5.1 Models

We conducted fraud detection using the following six machine learning models as shown in Table 4: Logistic Regression, Random Forest, Gradient Boost, CNN, Naïve Baysian, LightGBM and BERT.

Table 4. Description of Machine Learning Models for Job Posting Fraud Detection

Model	Description
Logistic Regression	Logistic regression is an appropriate regression model for binary classification problems. Logistic regression predicts the probability of outcome and classifies the data into different classes base on the pre-set probability threshold. Features generated from Topic modeling were input features for model training.
Random Forest	Random Forest is a bagging ensemble method that builds multiple uncorrelated trees. Each tree gives an independent prediction of the outcome. The outcome with the highest number of combined counts from the trees determines the model's prediction. Random forest has good predictive performance with reduced variance and bias. Features generated from Topic modeling were input features for model training. Optimal model is identified with the following hyperparameters: maximum_features=4, min_samples_leaf=3, maximum_depth=50, min_sample_split=8, n_estimators=1000..
Gradient Boost	Gradient Boost is a boosting ensemble classifier that builds simple decision trees sequentially, with each tree built to predict the residual error from the previous tree. Gradient Boost is useful for weak learners classification and the algorithm aims to minimise the loss functions. Features generated from Topic modeling were input features for model training. Hyperparameters of Gradient Boost were tuned to identify models that maximise the recall. Optimal models are identified with the following hyperparameters: learning_rate=0.5, n_estimators=80, minimum_sample_splits=50, maximum_depth=15, maximum_features=sqrt.
Convolutional Neural Network	In Convolutional Neural Network (CNN), filters of different matrix size were applied on the word vectors and output to convolved vectors. Max pooling of convolved vectors reduces the feature dimensions and prevent overfitting. Classification is performed at the fully connected layer which receives inputs from convolutional and pooling layers. CNN is a black box model with lowest interpretability but can achieve high prediction accuracy. The text to word sequencing prepares input vectors for CNN model.
Naïve Bayes	Naïve Bayes Classifier is based on Bayes Theorem for calculating probabilities and conditional probabilities to predict class of unknown data set. Classifier assumes that presence of particular feature in a class is independent to the presence of any other features. Features generated from Topic modeling were input features for model training.
LightGBM	LGBM is an implementation of Gradient Boosting Decision Tree algorithms with combination for gradient based on side sampling (GOSS) and exclusive feature binding (EFB) techniques. GOSS excludes data with smaller gradients, preferring instances with larger gradients for calculating information gain. EFB puts mutually exclusive features to reduce number of features without compromising accuracy of split point. LGBM has faster training speed, parallel learning support and low memory utilization. GridSearchCV provided following tuned hyperparameters: learning_rate=0.5, max_depth=20, metric='binary_logloss', num_leaves=90, objective='binary'
BERT	Bidirectional Encoder Representations from Transformers (BERT) is designed to pre-train deep bidirectional representations from unlabeled text. BERT's key innovation is applying bidirectional training of Transformer to language modelling. Transformer attention mechanism learns contextual relations between words based on all its surroundings text. Bidirectional training uses Masked Language Modelling (MLM) technique. Pretrained classification model on large corpus of unlabeled text including entire Wikipedia is fine-tuned with additional Binary Classification output layer.

5.2 Metrics

To evaluate the performance of our approach, we used accuracy as one of the metrics as it is commonly used to express rates of correct predictions; and the second metric to be used is the F1-measure, which is calculated from precision and recall:

$$\text{Accuracy} = \frac{a+b}{a+b+c+d}$$

$$\text{precision} = \frac{a}{(a+c)}$$

$$recall = \frac{a}{(a+d)}$$

$$\text{FM} = \frac{2a}{2a+c+d}$$

a: refers to the number of true positives (i.e. job listings that are correctly predicted "1")
b: refers to the number of true negatives (i.e. job listings that are correctly predicted "0")
c: is the number of false positives; and
d: is the number of false negatives.

In context to our use case, accuracy will measure the percentage of correct labels detected for fraudulent (fraudulent=1) and non-fraudulent jobs (fraudulent=0). accuracy measure based on this imbalanced data set will mislead our interpretation of result, hence techniques such as smote was applied.

The Area Under ROC Curve (AUC) quantifies the ability of classifiers to distinguish between fraudulent and non-fraudulent labels. The higher the AUC, the better our model is performing in distinguishing the positive and negative class labels. We focused on improving recall wherein it is acceptable for our system to flag non-fraudulent jobs mistakenly as fraudulent.

Models based on balanced train dataset were then evaluated using Recall, implying that the cost for wrongly classifying a fraudulent job as non-fraudulent is high and we should aim at increasing classification of True Positives. F1-measure is a combined measurement of precision and recall, which helps in striking a balance between classifying the True Negatives and restricting the system from over-sensitivity for fraudulent class label.

5.3 Models Performance

After applying SMOTE technique to overcome the class imbalance (Non-Fraudulent: 95%; Fraudulent:5%), the TRP scores (Key Metric) were higher for than the models' performance on the untreated imbalanced data.

Based on the plot of our ROC curves for the various models (Figure 5), Random Forest is suggested to be the best choice for our problem. We also observe the slight improvement in ROC values from SMOTE application as compared to the untreated data (Figure 6).

We also observe Model Performances from CNN and BERT are generally high despite not using the topic modelling features. Our suspicion is that there is information loss during data processing for topic modelling. However, we did not choose CNN and BERT as final best model as both models predict on text features. Other features which give significance to model prediction were not used by these models.

Figure 5. Combined ROC curves after SMOTE-treatment (Balanced Training Data)

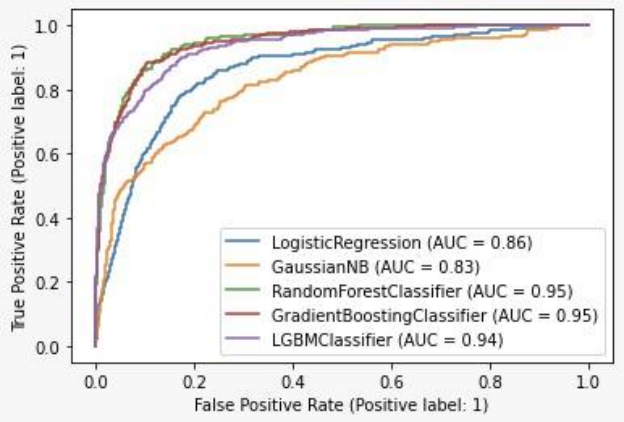


Figure 6. Combined ROC curves on Imbalanced Data

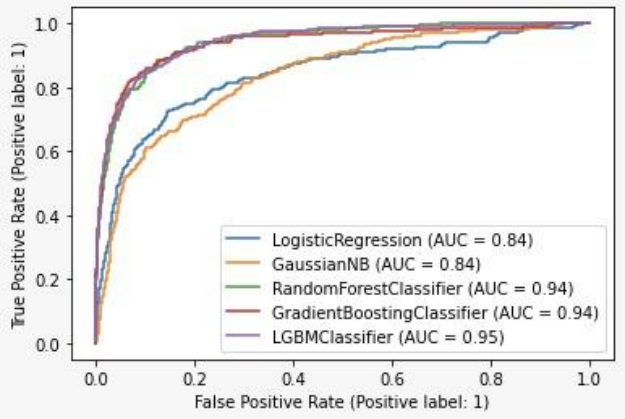


Table 5. Summary of Model Performance Evaluated on Testing Data

Metric	Naïve Bayes	Logistics	Random Forest	Gradient Boost	Light GBM	CNN	BERT
Imbalanced Training Data							
Accuracy	0.810	0.953	0.960	0.965	0.966	0.987	0.986
Precision	0.161	0.643	0.959	0.836	0.784	0.952	0.886
TPR	0.692	0.0692	0.181	0.354	0.404	0.765	0.809
TNR	0.816	0.998	0.999	0.996	0.994	0.998	0.994
AUC	0.84	0.84	0.94	0.94	0.94	0.98	0.902
SMOTE-treated Balanced Training Data							
Accuracy	0.747	0.831	**0.956**	0.961	0.956	NA	NA
Precision	0.131	0.191	**0.542**	0.609	0.544	NA	NA
TPR	0.75	0.769	**0.619**	0.569	0.6	NA	NA
TNR	0.747	0.834	**0.973**	0.981	0.974	NA	NA
AUC	0.83	0.86	**0.95**	0.95	0.94	NA	NA

5.4 Results Interpretation

Here we would like to share our findings on the 3 research questions set out at the start of our project.

RQ1: Which features are most associated with fraudulence?

After generating a feature importance ranking using the Random Forest Model (see Figure 7), we can see that a majority of the top 10 features are generated from topic modelling. An interesting observation is that the top two features in fraud prediction are 1) Having company logo, and 2) Missing a company profile.

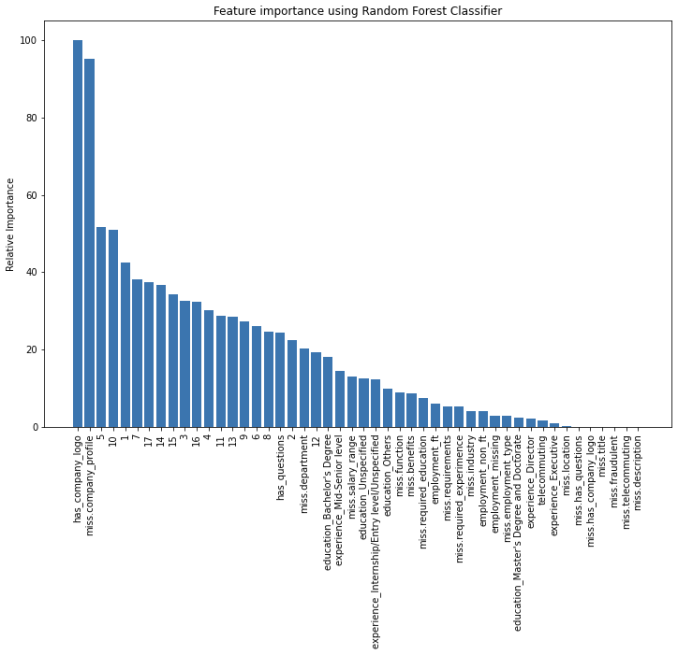
RQ2: What are the topics present in the company profile, job description, job requirements and job benefits?

In general, the topics were associated with the sector, the company or the skills of the corresponding job advertisement.

RQ3: Which ML model is most suitable to predict fraudulence?

The SMOTE treatment helped to improve the TPR scores across all models. The best performing models by AUC score based on the SMOTE-treated Balanced training data are 1) Random Forest and 2) Gradient Boost. We further decide Random Forest to be the most suitable model as it has a better TPR than Gradient Boost.

Figure 7. Feature Importance Ranking using Random Forest Model



Conclusion

5.5 Business Application

By deploying our selected model, job portals will be able to detect potential fraudulent job postings and prevent any applications to be received till the job posters provide additional supporting information.

Business value to the Job Portal

To convince Job Portals to utilize these models, we employ the Expected Value Framework (Sukup, 2019) to quantify the business impact they can expect to achieve. The formula used is as follows, and in essence a multiplication of values between "Cost Benefit Matrix" and Probability Matrix". The Probability Matrix is derived from the confusion matrix (see Figure 8) of our selected model, Random Forest.

E[X] = P(p) * [P(TP|p) * V(TP,p) + P(FN|p) * V(FN,p)] + P(n) * [P(FP|n) * V(FP,n) + P(TN|n) * V(TN,n)]

P: refers to probability of observing the class

V: refers to the associated value of the observed class

TP: refers to the Class "True Positive"

FN: refers to the Class "True Negative"

p: refers to Actual Positive observations

n: refers to Actual Negative observations

Figure 8. Confusion Matrix of Random Forest

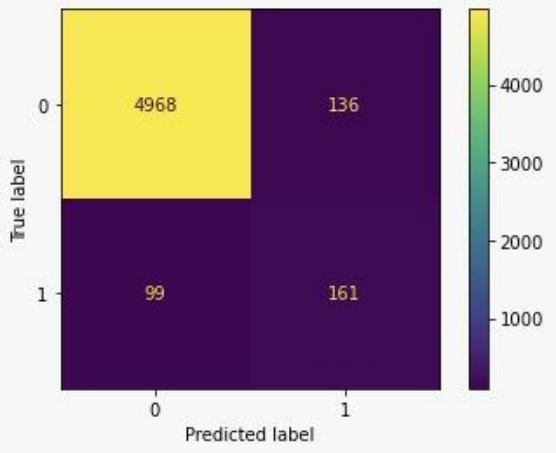


Table 6. Cost Benefit Matrix

	Actual Fraud	***Actual Non Fraud***
Predicted Fraud	$9,878[2]	-$267[1]
Predicted Non Fraud	-$9,878[2]	$0

[1] Requires 2 days of work from a recruiter, with a monthly salary of S$4,000

[2] Based off latest news report (Lim & Sun, 2022)

Table 7. Probability Matrix (based on RF Confusion Matrix)

	Actual Fraud	***Actual Non Fraud***
Predicted Fraud	TP (161/260) = 0.62	FP (136/5104) = 0.03
Predicted Non Fraud	FN (99/260) = 0.38	TN (4968/5104) = 0.97
Total	P (260/5364) = 0.048	N (5104/5364) = 0.952

Expected Value from our Anti Job Fraud Modelling

E[X] = 0.048 * [0.62 * $9,878 + 0.38 * -$9,878] + 0.952 * [0.03 * -$267 + 0.97 * $0] = Positive benefit of $106.17 per potential fraudulent job posting.

In summary, by implementing our model, the job portal can expect to receive an overall positive benefit equivalent to ~S$106 for every potential fraudulent posting identified.

5.6 Limitation of Current Study

Our existing dataset is based on the US market and may not accurately reflect the characteristics of the local (i.e. Singapore) market due to the lack of data from incumbent job portals.

Our models and packages are currently run based off text data from job postings listed in English. We may not achieve the same performance while applying to jobs postings of other languages.

5.7 Possible Future Work

We can improve the performance of our models by aggregating datasets across multiple job portals of the target local market by attuning to the characteristics of each region.

References

Janofsky, A. (2022, February 2). Scammers continue to spoof job listings to steal money and data, FBI warns. The Record by Recorded Future. https://therecord.media/scammers-continue-to-spoof-job-listings-to-steal-money-and-data-fbi-warns/

Lim, J., & Sun, D. (2022, January 26). Five types of job scams in S'pore and how to avoid them. The Straits Times. https://www.straitstimes.com/singapore/courts-crime/five-job-scam-variants

Linkedin. (2022). Community Report. Transparency Community Report. https://about.linkedin.com/transparency/community-report

Rafter, D. (2021, February 12). Job-posting scams and how to avoid them. NortonLifeLock. https://us.norton.com/internetsecurity-online-scams-job-posting-scams.html

Scamshield. (n.d.). ScamShield. https://www.scamshield.org.sg/

Sun, D., & Lim, J. (2022, January 26). $6.5m lost in 6 months: More falling prey to job scams. The Straits Times. https://www.straitstimes.com/singapore/courts-crime/job-scams-the-next-big-worry-says-head-of-anti-scam-centre

Sukup, J. (2019, November 14). Explain the "So What?" Behind Machine Learning Models with the Expected Value Framework (Part 2 of 3). Oracle AI & Data Science Blog. https://blogs.oracle.com/ai-and-datascience/post/explain-the-quotso-whatquot-behind-machine-learning-models-with-the-expected-value-framework-part-2-of-3

Codes and Data

GitHub - lttlrain/5153_Group_Project: 5153 Group project - Job Scam