The Hong Kong University of Science and Technology

SOSC 4300 Final Project

Fraudulent Job Posting detection using Computational Methods

1. Introduction

Society's increasing reliance on Internet access to fulfill personal tasks is giving rise to rampant internet fraud. In recent years, scammers have started searching for opportunities for job seekers. Online advertisement websites are the most common platform for recruitment and job searching. Yet, there is an increasing number of employment scams on the internet (Dutta & Bandyopadhyay, 2020). By posting fraudulent jobs through scam emails, scam URLs, or messages, recruitment fraud intends to grab money, personal information from job-seekers, or and steal their personal identities. Sometimes, it is difficult for job seekers to recognize whether a job post is real or fake. In view of this issue, we propose establishing a reliable and highly accurate model to detect fraudulent job posts. The goal of this project is to pave the way for the development of machine learning techniques to avoid job fraud on the Internet. We use the data from the Employment Scam Aegean Dataset (EMSCAD) as the information of online job advertisement and apply machine learning-based classification techniques to visualize topic modeling and to model the scam job advertisement detection. Four text analytics methods, which are Logistic Regression, Multinomial Naive Bayes Classifier, Support-Vector Machines Classifier (SVC), and Random Forest Classifier are used as classifiers for checking fraudulent job posts. After generating this data analytics, we will identify the best employment scam detection model by comparing their accuracy.

2. Background

In the era of growing social media and the internet, there is an increasing platform for employment recruitment for companies to reduce the extra cost of manpower, promotion during hiring progress. Due to the advantage of using online recruitment, there is an increasing number of online recruiting, as a result of the number of cybercrime cases — employment scams have become easier for the public to encounter. According to a cybersecurity ventures report, it predicts the annual cost of cybercrime damages over the world is estimated at around $6 trillion. Also, over 600 million jobs will be created in 2030 corresponding to the growth in job generation, referring to 2018 World bank statistics. Under the large demand in the job market, it is inevitable that employment scams are getting more serious in the next few years, even after post-COVID19. Most importantly, the most critical issues in the hiring process include collecting candidate's information, online assessment, employer's website, job portals. As employment scams are considered as Online recruitment Fraud (ORF), there are several research studies that have proven the hiring progress of recruitment fraud could not only harm the job-seekers in terms of privacy and taking money from them (Aurecon, n.d.) but also violate the credibility of the reputed company due to false representation (Dutta & Bandyopadhyay, 2020). In recent years, machine learning has been widely applied in fact-checking, such as Facebook, Youtube, Twitter using machine learning algorithms to detect sensitive content or false content. By considering the damages brought by the ORF, this study has modified and improved the existing classification algorithms in the previous research to better recognize the fake-job online posting, so as to hinder the growing number of employment scams.

3. Objectives

The main purpose of the paper is to find the best classification algorithm used for detecting online recruitment fraud. Specific objectives are as followed.

A. Enhance the accuracy of the model through pre-processing data
B. Apply feature selection techniques that assist to reduce the dimensionality
C. Build a reliable model to detect ads with the highest accuracy
D. Determine the best algorithm to predict the fraudulent online job advertisement

4. Data

In this study, we adopted the Employment Scam Aegean Dataset (EMSCAD) for detecting fake job postings. The Employment Scam Aegean Dataset was published by the University of the Aegean in 2016. The dataset is accessible to the public and consists of 17,880 real-world online job advertisements published between 2012 to 2014. Among the 17,880 online job advertisements, 17,014 of them are legalized and 886 of them are fraudulent (University of the Aegean, 2016). The dataset contains 18 columns that recorded the basic information of the job advertisements, such as its title, location of the job, salary range, company profile, and description of the job, etc. The legal and scam job advertisements are recognized in the column "fraudulent" by categorization (legalized job advertisement = 0; fraudulent job advertisement = 1). Data format includes string, HTML fragment, binary numbers, and nominal numbers.

Since there are only 886 observations that are fraudulent among the 17,880 online job advertisements, the dataset is severely imbalanced with an event rate of 4.96%. The legalized online job advertisement is the majority class while the fraudulent online job advertisement is the minority class in the predictive modeling. Inherently imbalanced data are in some common classification problems, such as fraud detection, spam detection, claim prediction, churn

prediction (Brownlee, 2020).

5. Proposed Methods

5.1 Pre-processing and data cleaning

Before selection and classification adoptions, there are several pre-processing techniques that will be applied to the dataset to ensure the dataset becomes more understandable. Pre-processing and data cleaning methods for the 17,880 online job advertisements can be identified into few parts: (1) Removal of unrelated data (2)Tokenization (3)Stemming (4)Veratization of text. First, Removal of unrelated data includes the elimination of missing values, stop words, extra space, irrelevant attributes. Second, tokenization refers to creating a token for the words obtained from the dataset, in other words, to divide the textual information into individual words. After that, Stemming is to extract the base-word form by reducing the affixes from them. Since there are different forms of the same English words, such as 'working', 'works', 'worked' from the base form "work", this technique could minimize the large variety of the word in the dataset. At last, after labeling the attributes and splitting the data into training and testing, the text will be vectorized by converting the text into numeral representation. In this study, we use the tf-idf model to identify the importance of those words based on their frequency in the text. Applying the above pre-processing technique can remove the tags, HTML, and the unnecessary columns, it could reduce the dimensional space of the documental matrix and the time for training the model (Agarwal, 2015).

5.2 Algorithms

A. Logistic Regression

Logistic regression is a supervised learning algorithm for classification in which the algorithm's outputs are categorized into one of the pre-chosen categories (Kambria, 2019). It examines the relationship between a dependent dichotomous variable and independent variables, in which the dependent variable follows Bernoulli Distribution (Navlani, 2019). The model seeks to predict the output value by locating the observations into the right category based on various input variables. The probability of the event occurrence will be computed in logistic regression. In our prediction modeling, the outcome is binary which is classified as one of the two classes "legalized" and "fraudulent". The algorithm will compute the occurrence probability that the online job advertisement is fraudulent. The threshold of the logistic regression model is set to 0.5. That is, any online job advertisements with a 50% or the higher probability of being fake will be predicted as "fraudulent" and any online job advertisements with a probability of less than 50% will be predicted as "legalized". Logistic regression is one of the most simple algorithms for classification with dichotomous outcomes which can be acted as a baseline for other classification (Navlani, 2019). In this study, we will build the logistic regression model in Scikit-learn implementation from Python.

B. SVC

To perform classification tasks, we use the support vector classifier class, which is written as 'SVC' and from the SVM library through Scikit-Learn implementation, which contains various algorithms for SVM. This method has specific parameters for different cases, such as simple SVM using 'linear' as value for linearly separable data and kernel SVM using Gaussian 'rbf', polynomial 'poly', sigmoid 'sigmoid', or computable kernel for non-linearly separable data lower dimensions to linearly separable data in higher dimension space (Yang, Li, & Yang, 2015).

In our study, we have considered some data that are not linearly separable and that it is hard to be recognized in linear/logistic regression and use different interaction of X to perfectly classify Y. Also, the Gaussian kernel can perform a perfect 100% prediction rate amongst the above kernel model while the other models have the issues of misclassified instances and are only suitable for binary classification. Therefore, we choose to use the Gaussian kernel and consider the Radial basis function kernel (rbf) as our value to explicitly project the data onto a higher-dimensional in a powerful way. Besides, there are few advantages of using Kernel SVM including minimizing the least square error, maximizing the margin under the best hyperplane, as well as accurately scattering.

C. Random Forest Classifier

We optimize a random forest model for data training from the scikit-learn implementation. In the random forest classifier, a multitude of decision trees are constructed through the technique of bagging, which is known as bootstrap aggregation, so there is random variation between trees in regards to the selected sample at each stage. In this model, the accuracy is higher as the higher variation of the decision trees in the random forest, and the better result delivered by an ensemble tree than an individual tree (Chen, 2006). In our implementation, all classes have weight one, the minimum number of samples required to split an internal node is 2 and the maximum depth of trees is none, therefore nodes are expanded until all leaves contain less than 2 samples. We set to allow bootstrap aggregation that the samples are drawn with replacement in order to grow trees with variation so that both the variance and the bias of the classification are reduced. Gini impurity is applied to measure the purity of the sub split and to select the best splitter when building an appropriate decision tree due to its higher efficiency compared with

entropy. To better the performance of the model, we obtain 100 classifying decision trees in the forest then the tendency of overfitting decreases and generalization error minimizes.

D. Multinomial Naive Bayes

Multinomial Naive Bayes classifier performs classification in describing the probability of counts in a number of categories, for example, it works for word frequency counts in text classification. The attributes of Multinomial Naive Bayes are multivariable observations with more than two classes in the model. It assumes that all attributes are independent given the context of the class. Given the attribute independence assumption, it avoids structure learning so as to simplify its parameter learning, especially with a large number of attributes. Therefore, multinomial naive Bayes is commonly used due to its simplicity, efficiency, and efficacy (Jiang, Wang, Li & Zhang, 2016). In our works, the additive Lidstone smoothing parameter is set to be 0.5 to adjust rare words occurring in the dataset so that their probabilities will not be exactly zero.

5.3 Evaluation parameters

A. Accuracy

Among the following evaluation parameters, the performance can be shown in the confusion matrix. For these metrics, accuracy is more straightforward to predict binary outcomes. The prediction ratio for the above method is the total number of predictions over the number of correct predictions. This overall ratio focused on True Positive and True Negative, while it neglects falsely classified positive and falsely classified negative values, so the result would be comparably high. Considering false positives and false negative cases, Precision and Recall can

help us to compensate for the shortcoming in accuracy. The formulas of accuracy are stated below.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$= \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + True\ Negative(TN) + False\ Positive(FP) + False\ Negative(FN)}$$

B. Precision

The Precision identifies the positive result from the true positive and false positive. This can be demonstrated by the ratio of the true positive result over the all positive result that predicted by the classifier. Since our data is imbalanced, this metric only considers the positive cases which leads to a biased result on positive cases, in the result of having higher prediction values. The formulas of precision are stated below.

$$Precision = \frac{True\ Positive\ (TP)}{Actual\ Results} \text{ or } \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

C. Recall

As opposed to the Precision metric, the recall will consider both the positive result and false negative. The ratio can represent true positives over all relevant results. Even if it is considered falsely classified negative, it tends to check all the positive classes and the number of positive classes, resulting in having the same problems with the Precision metric when producing the predicted result that is mostly biased towards the positive class, and also generate a high predictive value similar to accuracy & precision metrics. The formulas of the recall are stated below.

$$Recall = \frac{True\ Positive\ (TP)}{Predicted\ Results} \text{ or } \frac{True\ Positive\ (TP)}{True\ Positive + False\ Negative}$$

D. F-1 score

Considering the shortcoming of the above metrics, the f-1 score has taken precision and recall into account to ultimately measure the accuracy of the model. Since the above metrics have not specifically focused on the importance of falsely classified positive and falsely classified negative, the f-1 score has emphasized this value by giving more weighting for the false values, so as to prevent large amounts of true negative values affecting the accuracy score. The formula for the f-1 score is stated below.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

E. The ROC Curve and ROC AUC

The receiver operating characteristic curve (ROC curve) compiles the performance of a binary classification model on their positive class (Brownlee, 2020). It is plotted by the cumulative distribution function of True Positive in the y-axis against the cumulative distribution function of False Positive in the x-axis. The higher the true positive rate is, the better the model performs, whereas the better performance model is determined by the lower false-positive rate. The equations of true positive rate and false positive rate are stated below.

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive\ (FP)}{False\ Positive\ (FP) + True\ Negative\ (TN)}$$

The area under the receiver operating characteristic curve (ROC AUC) is a single-valued metric for calculating the probability of a classifier distinguishing true positive samples versus false-positive samples. When a model performs better in determining positive and negative instances between classes, the ROC curve approaches the coordinate (0,1) and AUC becomes greater.

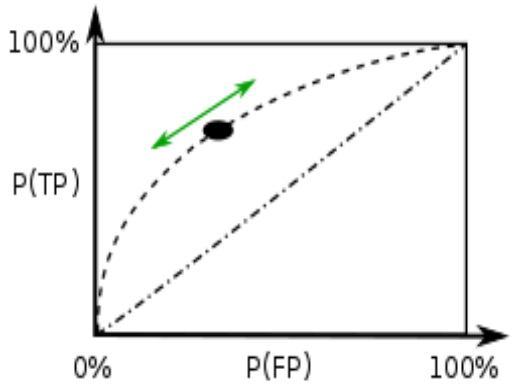

Figure 1 An example of ROC curves for prediction model

However, under the situation of an imbalanced dataset, where the low sample size of minority class (i.e. fraudulent job post), the false positive rate is weakened, and the ROC curve is over-optimized. ROC curve may not be the most reliable parameter to evaluate the performance in this case.

F. Precision-Recall Curve and Precision-Recall AUC

The precision-recall curve is the tradeoff between precision and recall (Zhang, 2020). Unlike the ROC curve, the precision-recall curve focuses on minority class (i.e. fraudulent job posts) and makes it a more effective parameter for evaluating the performance of different predictive classifier models under an imbalanced dataset. The precision-recall curve is generated by plotting precision metric on the y-axis versus recall metric on the x-axis. A model with perfect prediction performance is depicted at the point at the coordinate of (1,1), and thus the curve

representing a skillful model bow toward the point of the coordinate (1,1), while a no-skill model is a horizontal line at the bottom of the plot.

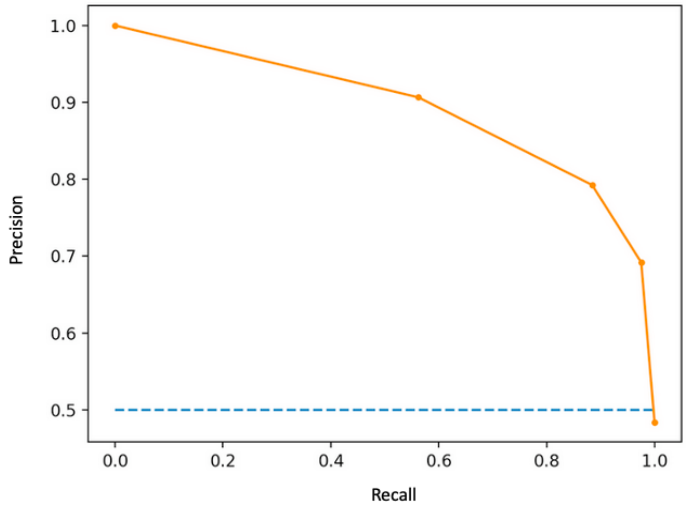

Figure 2. An example of Precision-recall curves.

6. Results

Table 1 shows a summary of the four classification models, including (1) Logistic Regression, (2) SVC, (3) Random Forest, (4) Multinomial Naive Bayes. By using logistic regression, 96% of the prediction in this is correct and the model is able to predict 96% of positive cases among the 17,880 online job advertisements. While the best score of the f-1 score is 1.0 and the worst is 0.0, the f-1 score of prediction using the logistic regression model is 0.95. For the SVC model, 97% of the prediction detected by this model is correct, 97% of positive cases is predicted, and the f-1 score for the SVC model is 0.97. For the random forest model, 98% of the prediction is correct, 97% of positive cases are correctly identified, and the f-1 score is 0.97. Lastly, for the multinomial Naïve Bayes model, 95% of the prediction is correct, 95% of positive cases is predicted in this model, and the f-1 score is 0.93.

Overall, the prediction of the random forest model obtained the highest scores in accuracy, precision, recall, and the f-1 score while the prediction of SVC obtained very similar results with random forest. The accuracy, precision, recall, and f-1 score of the prediction using multinomial Naïve Bayes and logistic regression models are slightly worse than the SVC and random forest model. Yet, four models obtained very high scores in accuracy, precision, recall, and f-1 score.

Table 1. Summary of Classification Report

Model	**Accuracy**	**Precision**	**Recall**	**F-1 score**
Logistic Regression	0.96	0.96	0.96	0.95
SVC	0.97	0.97	0.97	0.97
Random Forest	0.97	0.98	0.97	0.97
Multinomial Naive Bayes	0.95	0.95	0.95	0.93

Figure 3 shows the receiver operator characteristic (ROC) curves and the Area Under the Curve (AUC) of logistic regression, SVC, random forest classification, and the multinomial Naive Bayes classifier. The ROC curve is a performance measurement of classification which shows the diagnostic ability of each classifier and the trade-off between sensitivity and specificity (Chan, n.d.). The closer the curve is to the top-left corner, the better the performance of the classifier is. The ROC AUC is the area under the receiver operator characteristic curve which indicates how well the classifier is capable of distinguishing between classes and can be used for comparing different classifiers (Narkhede, 2018). The score of ROC AUC is between 0 to 1. The higher the ROC AUC score, the better the prediction performance of the classifier is (Narkhede, 2018). From Figure 3, the ROC curve of the SVC model is the closest to the top-left corner. The

ROC AUC scores of logistic regression, SVC, random forest classifier, and multinomial Naive Bayes model are 0.90, 0.91, 0.87, and 0.88 respectively. Among the ROC AUC of the four models, the ROC AUC of the SVC model obtains the highest score.

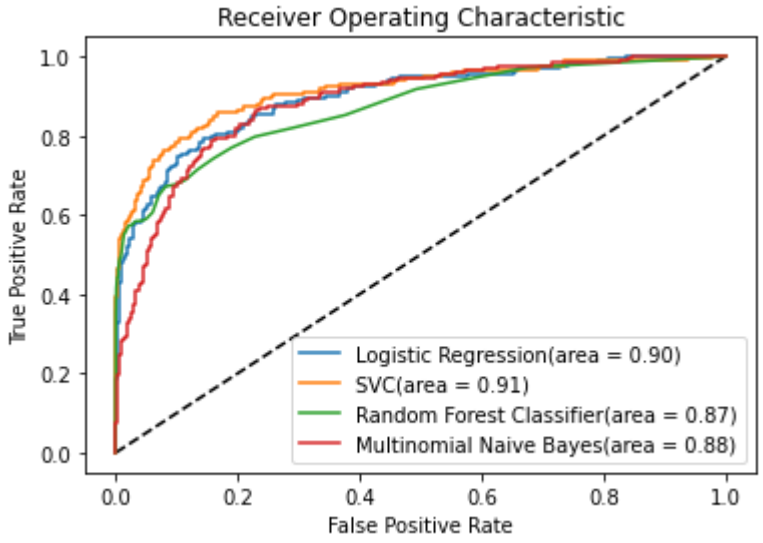

Figure 3. The ROC curves of the logistic regression, SVC, random forest, and multinomial Naive Bayes prediction

Figure 4 shows the precision-recall curve of the logistic regression model, SVC model, random forest classifier, and the multinomial Naive Bayes classifiers. Precision-Recall curves represent the tradeoff between precision and recall of the model for different thresholds. The higher and closer the curve towards the top-right corner, the more skillful is the model, and the higher the recall and precision of the model are. The higher the score of Area Under the Precision-Recall Curve (Precision-Recall AUC) for the precision-recall curve is, the better the predictions of the model are. The Precision-Recall AUC scores of logistic regression, SVC, random forest classifier, and multinomial Naive Bayes are 0.691, 0.837, 0.788, and 0.603 respectively.

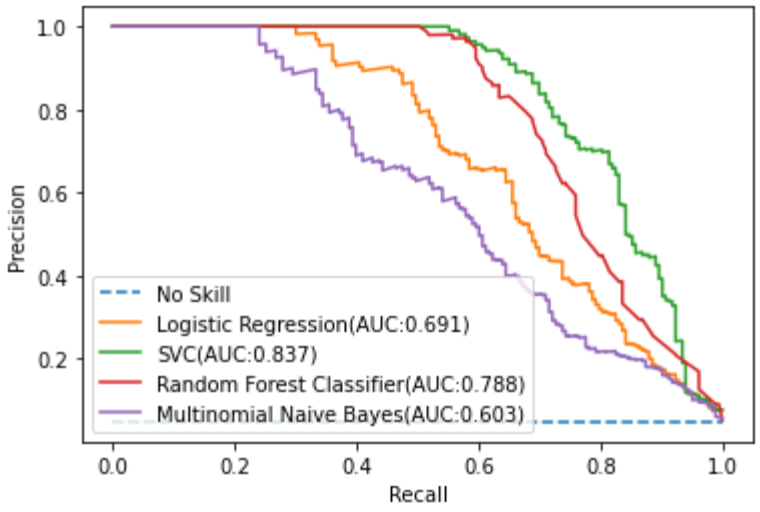

Figure 4. The precision-recall curves of the logistic regression model, SVC model, random forest classifier, and multinomial Naive Bayes classifier

Comparison

Since we used imbalanced datasets in this study, the number of legalized job advertisements is far more than the number of fraudulent job advertisements in prediction. When the dataset is imbalanced, the estimation of the ROC curve and ROC AUC score might be misleading and unreliable which easily achieves high accuracy in prediction. Since the ROC curve mainly focuses on predicting the majority class (legalized job posts are in our model) and leading to an over-optimistic evaluation of performance (Branco, Torgo, & Ribeiro, 2015). Therefore, to determine the best algorithm for predicting the fraudulent online job advertisement among the logistic regression, ROC curve and ROC AUC might not be falsifiable and not the best parameter to evaluate the performance of the model. Instead, the Precision-recall curve is recommended for evaluating the performance of different classifier models with highly skewed datasets (Branco, Torgo, & Ribeiro, 2015).

From the result, it shows that SVC, random forest classifier and multinomial Naive Bayes classifier, the ROC curves, the ROC AUC in Figure 3, and results from the classification table in Table 1, including the accuracy, precision, recall, and f-1 score are very close to each other. All four models obtained high scores in the above parameters. Hence, it is difficult to determine the best algorithm for detecting the fake job posting by solely looking at the classification table, ROC curves, and the scores of ROC AUC.

Yet, we can observe a significant difference in the result of the Precision-Recall curves and the Precision-Recall AUC curves in Figure 4. By comparing the scores of Precision-Recall AUC in Figure 4, we can conclude that the SVC model has the best prediction performance in detecting the fraudulent online job advertisement in our dataset, followed by random forest classifier, logistic regression, and the multinomial Naive Bayes classifier.

7. Conclusions

Throughout the whole project, we constructed four machine learning algorithms, including Logistic Regression, Multinomial Naive Bayes Classifier, Support-Vector Machines (SVM), and Random Forest Classifier, for fraudulent job posts detection. We have pre-processed the data by Removal of unrelated data, Tokenization, Stemming and Veratization of text in order to enhance the accuracy of each model and applied feature selection techniques to reduce the dimensionality. Besides, we used (1) Accuracy, (2) Precision, (3) Recall, (4) F-1 score, (5) ROC Curve and ROC AUC, and (6) Precision-Recall Curve and Precision-Recall AUC as the measurements of the reliability of every model in order to find out the highest accurate model in online job fraud detection. Since the dataset is imbalanced, the precision-recall curves generated

better prediction performance by showing significant differences than other parameters. The precision-recall curve focused on the minority class (fraudulent job post) while the ROC curve focused on the majority class (legalized job post), as a result, the precision-recall curve is the best parameter in comparing the performance of the four models. The result from the precision-recall curve demonstrated that the SVC model obtains the highest score which represents that it performs the best in identifying fraudulent job posts.

References

Agarwal, V. (2015). Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. *International Journal of Computer Applications,131*(4), 30-36. doi:10.5120/ijca2015907309

Alghamdi, B., & Alharby, F. (2019). An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security,10*(03), 155-176. doi:10.4236/jis.2019.103009

Branco, P., Torgo, L., & Ribeiro, R. (2015). A Survey of Predictive Modelling under Imbalanced Distribution. arXiv:1505.01658.

Brownlee, J. (2020). *ROC Curves and Precision-Recall Curves for Imbalanced Classification*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

Chan, C. (n.d.). *What is a ROC Curve and How to Interpret It. Displayr.* Retrieved from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Chen, C. H. (2006). *Signal and image processing for remote sensing*. CRC Press.

Dutta, S., & Bandyopadhyay, S. K. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. *International Journal of Engineering Trends and Technology,68*(4), 48-53. doi:10.14445/22315381/ijett-v68i4p209s

Freeze, D. (2020, November 19). Cybercrime To Cost The World $10.5 Trillion Annually By 2025. Retrieved from https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

Global Employment Trends 2012: World faces a 600 million jobs challenge, warns ILO. (2012, January 24). Retrieved from https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_171700/lang--en/index.htm

Jiang, L., Wang, S., Li, C., & Zhang, L. (2016). Structure extended multinomial naive Bayes. *Information Sciences*, *329*, 346-356.

Kambrai. (2019). Logistic Regression For Machine Learning and Classification. Retrieved from https://kambria.io/blog/logistic-regression-for-machine-learning/

Kohli, S. (2019). *Understanding a Classification Report For Your Machine Learning Model*. Retrieved from https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Laboratory of Information & Communication Systems Security. (2016). Employment Scam Aegean Dataset. University of the Aegean. Retrieved from http://emscad.samos.aegean.gr/

Narkhede, S. (2018). *Understanding AUC - ROC Curve*. Medium. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Navlani, A. (2019). Understanding Logistic Regression in Python. Datacamp. Retrieved from https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(1), 6. doi:http://dx.doi.org.lib.ezproxy.ust.hk/10.3390/fi9010006

Yang, Y., Li, J., & Yang, Y. (2015). The research of the fast SVM classifier method. *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. doi:10.1109/iccwamtip.2015.7493959