Executive Summary

Student Name: John C. Ludlum

Student ID: 001440662

Capstone Project Name: Predicting Fraudulent Job Advertisements using Logistic Regression

Problem Statement: A job advertisement scam is a fraudulent job advertisement typically found on the internet whose purpose is to steal money, obtain personal information, or harm the applicant in some way. Common tactics used by job advertisement scams include:

- Requiring applicants to pay fees to submit an application for a job that doesn't exist.
- Asking applicants to send resumés with personal information to untrustworthy sources.
- Tricking applicants into downloading malicious job application software containing computer viruses or malware.

Job advertisement scams cause many serious problems including financial loss, identity theft, and damaged reputations. Similar to the research carried out in (Vidros et al., 2017), (Alghamdi & Alharby, 2019) and (Kumar, 2020), the goal of the study was to create a model that predicts whether a job advertisement is fraudulent based on various advertisement features. In particular, logistic regression was used, a classic modeling technique for predicting the value of a binary target variable like fraudulent status (Patetta, Lesson 2.1).

Research Question: "Which features of a job advertisement can help identify whether the advertisement is fraudulent?"

Hypotheses: The following null hypothesis H_0 and alternative hypothesis H_a were used:

> H_0: There is no statistically significant association between the job advertisement features in the study and the probability of the advertisement being fraudulent.
>
> H_a: There is a statistically significant association between at least one of the job advertisement features in the study and the probability of the advertisement being fraudulent.

Data: The logistic regression model was developed using the Employment Scam Aegean Dataset (EMSCAD), a publicly available collection of 17,880 online job advertisements published between 2012 and 2014 which were collected by the University of the Aegean in Greece.[1] Of the 17,880 job advertisements in the dataset, 17,014 are classified as legitimate and 866 are classified as fraudulent.

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

The EMSCAD initially contained the following variables:

Table 1: List of variables in the original Employment Scam Aegean Dataset.

Variable Name	Type	Description
job_id	Numeric, discrete	ID number assigned to job advertisement. Equivalent to row number in dataset.
title	Character	Job title.
location	Character	Geographical location of job. Format: Country, Province, City.
department	Character	Internal department of company. (Ex: Marketing, Sales, etc.)
salary_range	Character	Lower & upper bounds for salary. (Ex: $20,000-$28,000)
company_profile	Text	Company profile.
description	Text	Job description.
requirements	Text	Eligibility requirements for job.
benefits	Text	List of job benefits.
telecommuting	Numeric, binary	True for telecommuting/work-from-home positions.
has_company_logo	Numeric, binary	True if company logo is visible in job advertisement.
has_questions	Numeric, binary	True if job advertisement has screening questions.
employment_type	Categorical	Employment type. (Ex: Full-time, part-time, etc.)
required_experience	Categorical	Prior experience required for job. (Ex: Entry level, executive, etc.)
required_education	Categorical	Education level required for job. (Ex: Bachelor's Degree, Master's Degree, etc.)
industry	Categorical	Industry company belongs to. (Ex: Telecommunications, financial services, etc.)
function	Categorical	Nature of job. (Ex: Sales, Engineering, etc.)
fraudulent	Numeric, binary	True if job advertisement is fraudulent.

Data Preparation: The following changes were made to the EMSCAD to prepare the data for logistic regression:

- The character variable *location* stating the country, province, and city of the job was replaced by a categorical variable called *country* which only states the country.
- Due to high sparsity (83.96%), the character variable *salary_range* was replaced by a binary variable *mentions_salary* which is true if salary is mentioned in the job advertisement or false if salary is not mentioned.

- All categories of the "Vocational" type for *required_education* were merged into a single category. Also, the categories "Some High School Coursework" and "High School or equivalent" were merged.
- Missing values for the categorical variables *employment_type*, *required_experience*, *required_education*, *industry*, *function*, and *country* were assigned the category "Unspecified."
- The character variable *department* was dropped because it had a high sparsity (64.61%) and was similar to the categorical variable *function* which is less sparse (36.10%).
- To avoid problems such as high-dimensionality and quasi-complete separation, the levels of the categorical variables *industry*, *function*, and *country* were collapsed. *country* was replaced by the binary variable *from_US* which is true if the job advertisement comes from the United States or false if it comes from any other country including "Unspecified." *industry* and *function* were collapsed using the smooth weight-of-evidence technique (SWOE) described in (Patetta, Lesson 3.2).
- The character variable *title* and the text variables *company_profile*, *description*, *requirements*, and *benefits* could not be used directly in the logistic regression model because they are not numeric or categorical (Patetta, Lesson 3.2). Instead, key information was extracted from them using Python's NLTK text mining library:
  - The binary variables *has_fraud_bigram* and *has_legit_bigram* were created which are true if the job advertisement contains a two-word phrase commonly found in fraudulent or legitimate job advertisements, respectively.
  - The binary variables *has_email*, *has_phone*, and *has_url* were created which are true if the job advertisement's text features contain an email address, phone number, or a link to an external website, respectively.
  - The binary variable *money_in_title* was created which is true if the character variable *title* contains a "$" symbol.
  - The number of words in the text features were calculated and stored as *company_profile_length*, *description_length*, *requirements_length*, and *benefits_length*.

The resulting EMSCAD_final dataset had the following variables:

Table 2: List of variables in the "EMSCAD_final" dataset.

Variable Name	Type	Description
employment_type	Categorical (6 levels)	Employment type. (Ex: Full-time, part-time, etc.)
required_experience	Categorical (8 levels)	Prior experience required for job. (Ex: Entry level, executive, etc.)
required_education	Categorical (10 levels)	Education level required for job. (Ex: Bachelor's Degree, Master's Degree, etc.)

company_profile_length	Numeric, discrete	Number of words in *company_profile*.
description_length	Numeric, discrete	Number of words in *description*.
requirements_length	Numeric, discrete	Number of words in *requirements*.
benefits_length	Numeric, discrete	Number of words in *benefits*.
industry_SWOE	Numeric, continuous	Smooth weight-of-evidence technique applied to *industry*.
function_SWOE	Numeric, continuous	Smooth weight-of-evidence technique applied to *function*.
from_US	Numeric, binary	True if job advertisement is from the United States.
money_in_title	Numeric, binary	True if *title* contains "$" symbol.
mentions_salary	Numeric, binary	True if job advertisement mentions salary.
telecommuting	Numeric, binary	True for telecommuting/work-from-home positions.
has_company_logo	Numeric, binary	True if company logo is visible in job advertisement.
has_questions	Numeric, binary	True if job advertisement has screening questions.
has_email	Numeric, binary	True if job advertisement contains email address.
has_phone	Numeric, binary	True if job advertisement contains phone number.
has_url	Numeric, binary	True if job advertisement contains link to external website.
has_fraud_bigram	Numeric, binary	True if job advertisement contains a bigram commonly found in fraudulent job advertisements (see pg. 25).
has_legit_bigram	Numeric, binary	True if job advertisement contains a bigram commonly found in legitimate job advertisements (see pg. 25).
fraudulent	Numeric, binary	True if job advertisement is fraudulent.

The first 20 variables are independent variables (either categorical or numeric) which can all be used directly to predict the value of *fraudulent*, the binary dependent variable that was the focus of the study.

Data Analysis: In the first stage of the analysis, several different types of charts and graphs were created to visualize the distribution of fraudulent job advertisements across the independent variables. These charts and graphs provided many useful insights into the distinguishing features of fraudulent job advertisements such as:

- Even though a majority (64.99%) of the job advertisements are full-time positions which have the most fraudulent cases (490) by sheer numbers, jobs where commitment level is low ("Part-time", "Other", "Unspecified") have the highest fraudulent rates (9.28%, 6.94%, 6.61%, respectively).
- The most common required experience level was "Unspecified" (39.43%) which also had the most fraudulent cases (435). However, jobs where required experience was either very high ("Executive") or very low ("Unspecified", "Entry level") had higher fraudulent rates (7.09%, 6.64%, 6.17%, respectively) than job advertisements with moderate required experience levels ("Associate", "Mid-Senior level").
- The most common required education category was also "Unspecified" (53.14%) which again had the most fraudulent cases (512). However, the "Certification" category had the highest fraudulent rate (11.18%) followed by "High School or equivalent" (9.02%). This makes sense because job advertisement scams are unlikely to require high education levels as this would discourage certain groups of people from applying and decrease the scam's chances of success.
- Fraudulent job advertisements are consistently shorter (lower average number of words) than legitimate job advertisements across all four text features (company profile, description, requirements, benefits).
- Job advertisements which mention money in the title are very rare (0.73% of data) but when they do occur, there is a high probability (44.62%) that they are fraudulent.
- When a job advertisement comes from the United States, mentions salary, supports telecommuting, or contains an email address or phone number, the probability that the advertisement is a scam increases. On the other hand, when the job advertisement has a company logo or screening question, the probability that the advertisement is a scam decreases.
- The presence of a link to an external website did not have a significant effect on fraudulent status. Advertisements with URLs had a fraudulent rate of 4.89% while advertisements which did not had fraudulent rate of 4.75%.
- Advertisements with a common fraud bigram have a higher fraudulent rate (14.58%) than those which do not (2.50%) while advertisements which have a common legitimate bigram have a lower fraudulent rate (3.89%) than those which do not (6.64%).

The second stage of the analysis was making the logistic regression model. The stepwise selection model using a 5% significance level as the selection criterion was found to be superior (82.56% accuracy, 15 predictors, AUC=0.9320) to the all-variables model (82.47% accuracy, 20 predictors, AUC=0.9255). Because there is at least one predictor variable in the stepwise selection model that is significant at the 5% significance level, we reject the null hypothesis H_0 in favor of the alternative hypothesis H_a. Table 3 shows the parameter estimates for the stepwise selection model.

Table 3: Analysis of maximum likelihood estimates (stepwise selection model).

Parameter	Category	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.1268	0.3928	63.3686	<.0001
required_education	Associate Degree	1	0.7135	0.7659	0.8677	0.3516
required_education	Bachelor's Degree	1	0.3599	0.2185	2.7129	0.0995
required_education	Certification	1	0.2138	0.6628	0.1041	0.7470
required_education	Doctorate	1	-11.8825	1693.7	0.0000	0.9944
required_education	High School or equivalent	1	0.2652	0.2143	1.5315	0.2159
required_education	Master's Degree	1	1.6393	0.4162	15.5145	<.0001
required_education	Professional	1	-1.0105	1.3830	0.5338	0.4650
required_education	Some college coursework	1	1.0298	0.9738	1.1184	0.2903
required_education	Vocational	1	-14.2418	661.8	0.0005	0.9828
company_profile_length		1	-0.00929	0.00165	31.6527	<.0001
description_length		1	-0.00103	0.000522	3.8814	0.0488
requirements_length		1	-0.00211	0.000903	5.4356	0.0197
benefits_length		1	0.00442	0.00144	9.4155	0.0022
industry_SWOE		1	0.8364	0.0771	117.6376	<.0001
function_SWOE		1	0.5856	0.0961	37.0939	<.0001
from_US		1	1.1151	0.1797	38.5176	<.0001
money_in_title		1	2.0096	0.4630	18.8402	<.0001
mentions_salary		1	0.9984	0.1887	28.0055	<.0001
has_company_logo		1	-1.5436	0.2004	59.3067	<.0001
has_email		1	0.7796	0.3174	6.0338	0.0140
has_phone		1	0.8618	0.3686	5.4669	0.0194
has_fraud_bigram		1	1.4027	0.1531	83.9056	<.0001
has_legit_bigram		1	-0.4563	0.1444	9.9812	0.0016

Of the 15 job advertisement features in the stepwise selection model, eight of them (*company_profile_length*, *industry_SWOE*, *function_SWOE*, *from_US*, *money_in_title*, *mentions_salary*, *has_company_logo*, and *has_fraud_bigram*) were highly significant predictors of fraudulent status with Wald Test p-values below 0.001. Also noteworthy were several of the odds ratio estimates:

- The odds ratio for *from_US* is 3.050 which means that the odds of a job advertisement being fraudulent are 3.050 times higher if the advertisement comes from the United States than if it comes from another country.
- The odds ratio for *has_fraud_bigram* is 4.066 which means that the odds of a job advertisement being fraudulent increase by a factor of 4.066 when the advertisement contains a common fraud bigram.
- The odds ratio for *money_in_title* is 7.460 which means that job advertisements which mention money in the title are 7.460 times more likely to be fraudulent than advertisements which do not.
- The odds ratio for *has_company_logo* is 0.214. From this odds ratio, we can infer that the odds of a job advertisement being fraudulent decrease by 78.6% when the job advertisement displays a company logo.
- The odds ratios for *company_profile_length*, *description_length*, and *requirements_length* are 0.991, 0.999, and 0.998, respectively. The fact that these odds ratios are below 1 implies that as the number of words in a job advertisement increases, the odds of the advertisement being fraudulent decrease.

Limitations of techniques and tools used: There were several limitations to the study such as:

- The only predictive modeling technique that was considered was logistic regression. Even though the stepwise selection model's 82.56% accuracy rate is quite good, there are many other predictive modeling techniques which could achieve better results. There might even be other logistic regression models besides the all-variables model and the stepwise selection model that were considered which perform better.
- The job advertisements in the EMSCAD are all written in English. There is no guarantee that the results of the study will generalize to job advertisements written in other languages, especially the text mining results.
- The job advertisements in the EMSCAD were published between 2012 and 2014. It is possible that fraudulent job advertisements today in the year 2021 have new characteristics which were not detected by the study.

Summary of proposed actions: To avoid becoming the victim of a job advertisement scam, the public can follow these six guidelines when researching and applying for jobs on the internet:

1. Be extremely cautious about job advertisements which mention money in the title. These advertisements were rarely encountered in the study, but when they did appear, they had an alarmingly high fraudulent rate of 44.62%. In the stepwise selection model, the binary variable *money_in_title* was one of the strongest predictors of fraudulent status. According to the model, a job advertisement which mentions money in the title is 7.460 times more likely to be a scam than advertisements which do not.
2. Be wary of job advertisements which are very short and provide little information about the position. The study found that on average, fraudulent job advertisements are shorter than legitimate job advertisements when it comes to text features such as company profile, description, requirements, and benefits. In the stepwise selection model, an increase in the number of words in the advertisement decreases the odds of the advertisement being fraudulent.
3. Watch out for advertisements which contain phrases that sound too good to be true such as "No experience required!", "Work from home!" or "Signing bonus available!" The study found that fraudulent job advertisements frequently contain these types of phrases. In the stepwise selection model, the binary variable *has_fraud_bigram* was a highly significant predictor of fraudulent status. According to the model, if a job advertisement contains a common fraud bigram, the odds of the advertisement being fraudulent increase by a factor of 4.066.
4. Look for signs that the advertisement comes from a real-life company. For example, the study found that job advertisements which have screening questions or display a company logo have lower fraudulent rates than advertisements which do not. In the stepwise selection model, the presence of a company logo decreases the odds of the advertisement being fraudulent by 78.6%.
5. Be cautious about advertisements which require little commitment, experience, or education. While many legitimate businesses do post advertisements for part-time or entry-level positions on the internet, the study found that these types of advertisements have higher fraudulent rates than advertisements for full-time positions requiring prior experience and college education. The categorical variable *required_education* was identified as a significant predictor of fraudulent status in the stepwise selection model.
6. Overall, be sure to exercise the same common sense when researching and applying for jobs online as with other internet activities. Do not give personal information to an online recruiter unless you are absolutely sure that they can be trusted. Fraudulent job advertisements contain email addresses and phone numbers more often than legitimate job advertisements. Be wary about paying a fee to submit an application—most companies do not impose them (Reinicke, 2020). Do not click on a link to an external website or download any job application software if it seems suspicious in any way.

Expected benefits of the study:

- The stepwise selection model can be reused to predict whether other job advertisements besides those in the EMSCAD are fraudulent, although further research is needed to determine how well the model generalizes to modern advertisements.
- The six guidelines for avoiding job advertisement scams can help the public avoid fraudulent job advertisements and the problems caused by them (financial loss, identity theft, damaged reputations, etc.)

References:

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

Kumar, Vaibhav. (2020, June 6). *Classifying Fake and Real Job Advertisements using Machine Learning*. Analytics India Magazine. https://analyticsindiamag.com/classifying-fake-and-real-job-advertisements-using-machine-learning/

Patetta, Mike. (n.d.). *Predictive Modeling Using Logistic Regression.* SAS Training Courses. https://support.sas.com/edu/schedules.html?crs=PMLR&ctry=US

Reinicke, Carmen. (2020, October 6). *Job scams have increased as COVID-19 put millions of Americans out of work*. CNBC. https://www.cnbc.com/2020/10/06/job-scams-have-increased-during-the-covid-19-crisis-how-to-one.html

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), pp. 6-25. https://doi.org/10.3390/fi9010006