PREDICTING FRAUDULENT JOB ADVERTISEMENTS

USING LOGISTIC REGRESSION

A Capstone Project

Submitted to the College of Information Technology

in Partial Fulfillment of the Requirements for the Degree

Master of Science in Data Analytics

John Ludlum

Western Governors University

June 3, 2021

ABSTRACT

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. There are serious consequences to falling victim to a job advertisement scam including financial loss, identity theft, and damaged reputations. In this study, a logistic regression model is constructed to predict whether a job advertisement is fraudulent based on various features of the advertisement. The model is developed using the Employment Scam Aegean Dataset, a publicly available dataset of 17,880 online job advertisements published between 2012 and 2014 that were classified as legitimate or fraudulent by researchers at the University of the Aegean. The final model identifies 15 job advertisement features as being statistically significant predictors of fraudulent status at the 5% significance level. In addition, a set of six guidelines to help the public avoid job advertisement scams is provided.

TABLE OF CONTENTS

Introduction 1

Context 1

Research Question 2

Hypotheses 2

Data Collection 3

The Employment Scam Aegean Dataset 3

List of Original Variables 3

Data Preparation and Extraction 5

Tools and Techniques 5

Exploring the Data 5

Cleaning the Data 13

Collapsing Categorical Variable Levels 17

Text Mining 19

List of New Variables 28

Analysis 31

Visualizations 31

Training and Validation Datasets 43

All-Variables Model 44

Stepwise Selection Model 50

Comparing the Models 54

Conclusion 57

Summary of Findings 57

Limitations 59

Further Research 60

References 61

Appendix 62

INTRODUCTION

**Context**

Now more than ever, finding a job is an online process. Within the United States, a 2015 survey by the Pew Research Center found that 54% of Americans have researched jobs on the internet and 45% have applied for a job online (Maurer, 2015). The number of Americans using the internet to research and apply for jobs continues to grow, especially during the ongoing COVID-19 pandemic when millions of Americans have lost their jobs and are seeking new employment (Reinicke, 2020).

Even though it is convenient to search for jobs online, one of the drawbacks is the threat of falling victim to a job advertisement scam. 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 but are not limited to:

* 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.

The problems caused by job advertisement scams (financial loss, identity theft, damaged reputations, etc.) are serious enough that researchers are working to identity their common attributes and educate the general public about how to avoid them (Vidros et al., 2017).

**Research Question**

The research question for the study is: “Which features of a job advertisement can help identify whether the advertisement is fraudulent?” Following in the footsteps of (Vidros et al., 2017), (Alghamdi & Alharby, 2019) and (Kumar, 2020), the goal of the study is to create a model that will predict whether a job advertisement is fraudulent based on various features of the advertisement. In particular, the study will use logistic regression, a classic modeling technique for predicting the value of a binary target variable like fraudulent status (Patetta, Lesson 2.1). Once the logistic regression model has been created, it will be possible to answer the research question by examining the most statistically significant predictor variables in the model. The study will be beneficial to the general public because it will provide insights into the common features of job advertisement scams and give guidance for how to avoid them.

**Hypotheses**

The following null hypothesis and alternative hypothesis will be used when constructing the logistic regression model:

: There is no statistically significant association between the job advertisement features in the study and the probability of the advertisement being fraudulent.

: 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.

It is expected that the null hypothesis will be rejected in favor of the alternative hypothesis because the model will incorporate a wide variety of job advertisement features and the probability that none of them are significant is low. However, which features specifically are significant remains to be determined.

DATA COLLECTION

**The Employment Scam Aegean Dataset**

The logistic regression model will be developed using the Employment Scam Aegean Dataset (EMSCAD), a collection of 17,880 online job advertisements published between 2012 to 2014 which were collected by the University of the Aegean in Greece. Of the 17,880 job advertisements in the dataset, 17,014 are classified as legitimate and 866 are classified as fraudulent. The dataset is publicly available and can be downloaded as a CSV file from the University of the Aegean website.[[1]](#footnote-1) The dataset has also been investigated by data scientists on Kaggle and can be downloaded there as well.[[2]](#footnote-2) No particular methodology was needed to collect the data since this step was already completed by the researchers at the University of the Aegean when they assembled the EMSCAD.

**List of Original 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. |

Variables 2-17 (*title*, *location*, *department*, *salary\_range*, *company\_profile*, *description*, *requirements*, *benefits*, *telecommuting*, *has\_company\_logo*, *has\_questions*, *employment\_type*, *required\_experience*, *required\_education*, *industry*, *function*) are independent variables that will be used to predict the value of Variable 18, *fraudulent*, the binary dependent variable that is the focus of the study. However, one of the challenges presented by the EMSCAD is that logistic regression models only accept numeric variables or categorical variables encoded with dummy variables as inputs (Patetta, Lesson 3.2), and some of the independent variables are character or text variables. To address this issue, the study will adopt a similar approach as (Vidros et al., 2017) and create new numeric variables based on the information provided by the character and text variables which can then be used in the model.

DATA PREPARATION AND EXTRACTION

**Tools and Techniques**

The majority of the analysis will be carried out using SAS OnDemand for Academics, a free cloud version of SAS Studio for university students and professors. SAS is a statistical software suite that has been a data science industry standard for decades (Goled, 2020). SAS is an appropriate choice of software for the study because it provides all of the necessary tools for processing the data, making graphs and charts, and creating logistic regression models. However, a small but significant portion of the analysis will be performed outside of SAS using the popular object-oriented programming language Python. In particular, the character variable *title* and the text variables *company\_profile*, *description*, *requirements*, and *benefits* will undergo text mining with Python’s Natural Language Tool Kit (NLTK) to create new numeric variables which can be used in the logistic regression model since the original character and text variables cannot be used directly themselves. SAS does have its own text mining software called SAS Text Miner, but it is not free and open-source like Python’s NLTK library.

**Exploring the Data**

In SAS Studio, the first step was to create a project library called “capstone” and import the EMSCAD CSV file as a SAS table:

/\* Create project library \*/

libname capstone "&path";

/\* Import data \*/

**proc** **import** datafile="&path/EMSCAD.csv"

dbms=csv

out=capstone.EMSCAD;

guessingrows=max;

**run**;

Next, the data sparsity was investigated using the following code from (Wicklin, 2011):

/\* Create missing/not missing format \*/

**proc** **format**;

value $missfmt ' '='Missing' other='Not Missing';

value missfmt . ='Missing' other='Not Missing';

**run**;

/\* View missing/not missing statistics for each variable \*/

**proc** **freq** data=capstone.EMSCAD;

format \_CHAR\_ $missfmt.;

tables \_CHAR\_ / missing missprint nocum;

format \_NUMERIC\_ missfmt.;

tables \_NUMERIC\_ / missing missprint nocum;

**run**;

/\* Overall data sparsity \*/

%let all\_vars=job\_id, title, location, department, salary\_range, company\_profile, description, requirements, benefits, telecommuting, has\_company\_logo, has\_questions, employment\_type, required\_experience, required\_education, industry, function, fraudulent;

**proc** **sql**;

select sum(cmiss(&all\_vars))/(**17880**\***18**)

as **'Overall sparsity'n**

from capstone.EMSCAD;

**quit**;

The results are summarized in Table 2.

**Table 2:** Data sparsity results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Missing**  **(Count)** | **Missing**  **(Percent)** | **Not Missing**  **(Count)** | **Not Missing**  **(Percent)** |
| job\_id | 0 | 0 | 17880 | 100 |
| title | 0 | 0 | 17880 | 100 |
| location | 346 | 1.94 | 17534 | 98.06 |
| department | 11553 | 64.61 | 6327 | 35.39 |
| salary\_range | 15012 | 83.96 | 2868 | 16.04 |
| company\_profile | 3308 | 18.50 | 14572 | 81.50 |
| description | 0 | 0 | 17880 | 100 |
| requirements | 2696 | 15.08 | 15184 | 84.92 |
| benefits | 7206 | 40.30 | 10674 | 59.70 |
| telecommuting | 0 | 0 | 17880 | 100 |
| has\_company\_logo | 0 | 0 | 17880 | 100 |
| has\_questions | 0 | 0 | 17880 | 100 |
| employment\_type | 3471 | 19.41 | 14409 | 80.59 |
| required\_experience | 7050 | 39.43 | 10830 | 60.57 |
| required\_education | 8105 | 45.33 | 9775 | 54.67 |
| industry | 4903 | 27.42 | 12977 | 72.58 |
| function | 6455 | 36.10 | 11425 | 63.90 |
| fraudulent | 0 | 0 | 17880 | 100 |
| **Overall sparsity:** 21.78% | | | | |

From Table 2, we can see that the variables *job\_id*, *title*, *description*, *telecommuting*, *has\_company\_logo*, *has\_questions* and *fraudulent* have no missing values, the variables *location*, *company\_profile*, *requirements*, *benefits*, *employment\_type*, *required\_experience*, *required\_education*, *industry*, and *function* have some missing values (below 50%), and the variables *department* and *salary\_range* have many missing values (at least 60%). Overall, the dataset is 78.22% occupied, 21.78% sparse. The missing values in the dataset will be addressed in the next section, Cleaning the Data.

Next, the categorical variables *employment\_type*, *required\_experience*, *required\_education*, *industry*, and *function* were explored in more detail:

/\* Inspect raw categorical variables \*/

%let cat\_vars=employment\_type required\_experience required\_education

industry function;

**proc** **freq** data=capstone.EMSCAD order=freq;

tables &cat\_vars;

**run**;

The results of the FREQ procedure are shown in Tables 3-7.

**Table 3:** PROC FREQ results for *employment\_type*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| Full-time | 11620 | 80.64 | 11620 | 80.64 |
| Contract | 1524 | 10.58 | 13144 | 91.22 |
| Part-time | 797 | 5.53 | 13941 | 96.75 |
| Temporary | 241 | 1.67 | 14182 | 98.42 |
| Other | 227 | 1.58 | 14409 | 100.00 |
| **Frequency Missing:** 3471 | | | | |

**Table 4:** PROC FREQ results for *required\_experience*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| Mid-Senior level | 3809 | 35.17 | 3809 | 35.17 |
| Entry level | 2697 | 24.90 | 6506 | 60.07 |
| Associate | 2297 | 21.21 | 8803 | 81.28 |
| Not Applicable | 1116 | 10.30 | 9919 | 91.59 |
| Director | 389 | 3.59 | 10308 | 95.18 |
| Internship | 381 | 3.52 | 10689 | 98.70 |
| Executive | 141 | 1.30 | 10830 | 100.00 |
| **Frequency Missing:** 7050 | | | | |

**Table 5:** PROC FREQ results for *required\_education*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| Bachelor's Degree | 5145 | 52.63 | 5145 | 52.63 |
| High School or equivalent | 2080 | 21.28 | 7225 | 73.91 |
| Unspecified | 1397 | 14.29 | 8622 | 88.20 |
| Master's Degree | 416 | 4.26 | 9038 | 92.46 |
| Associate Degree | 274 | 2.80 | 9312 | 95.26 |
| Certification | 170 | 1.74 | 9482 | 97.00 |
| Some College Coursework Completed | 102 | 1.04 | 9584 | 98.05 |
| Professional | 74 | 0.76 | 9658 | 98.80 |
| Vocational | 49 | 0.50 | 9707 | 99.30 |
| Some High School Coursework | 27 | 0.28 | 9734 | 99.58 |
| Doctorate | 26 | 0.27 | 9760 | 99.85 |
| Vocational - HS Diploma | 9 | 0.09 | 9769 | 99.94 |
| Vocational - Degree | 6 | 0.06 | 9775 | 100.00 |
| **Frequency Missing:** 8105 | | | | |

**Table 6:** PROC FREQ results for *industry*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| Information Technology and Services | 1734 | 13.36 | 1734 | 13.36 |
| Computer Software | 1376 | 10.60 | 3110 | 23.97 |
| Internet | 1062 | 8.18 | 4172 | 32.15 |
| Marketing and Advertising | 828 | 6.38 | 5000 | 38.53 |
| Education Management | 822 | 6.33 | 5822 | 44.86 |
| Financial Services | 779 | 6.00 | 6601 | 50.87 |
| Hospital & Health Care | 497 | 3.83 | 7098 | 54.70 |
| Consumer Services | 358 | 2.76 | 7456 | 57.46 |
| Telecommunications | 342 | 2.64 | 7798 | 60.09 |
| Oil & Energy | 287 | 2.21 | 8085 | 62.30 |
| Retail | 223 | 1.72 | 8308 | 64.02 |
| Real Estate | 175 | 1.35 | 8483 | 65.37 |
| Accounting | 159 | 1.23 | 8642 | 66.59 |
| Construction | 158 | 1.22 | 8800 | 67.81 |
| E-Learning | 139 | 1.07 | 8939 | 68.88 |
| Management Consulting | 130 | 1.00 | 9069 | 69.89 |
| Design | 129 | 0.99 | 9198 | 70.88 |
| Health, Wellness and Fitness | 127 | 0.98 | 9325 | 71.86 |
| Staffing and Recruiting | 127 | 0.98 | 9452 | 72.84 |
| Insurance | 123 | 0.95 | 9575 | 73.78 |
| Automotive | 120 | 0.92 | 9695 | 74.71 |
| Logistics and Supply Chain | 112 | 0.86 | 9807 | 75.57 |
| Human Resources | 108 | 0.83 | 9915 | 76.40 |
| Online Media | 101 | 0.78 | 10016 | 77.18 |
| Apparel & Fashion | 97 | 0.75 | 10113 | 77.93 |
| Legal Services | 97 | 0.75 | 10210 | 78.68 |
| Facilities Services | 94 | 0.72 | 10304 | 79.40 |
| Hospitality | 88 | 0.68 | 10392 | 80.08 |
| Computer Games | 86 | 0.66 | 10478 | 80.74 |
| Banking | 84 | 0.65 | 10562 | 81.39 |
| Building Materials | 78 | 0.60 | 10640 | 81.99 |
| Leisure, Travel & Tourism | 76 | 0.59 | 10716 | 82.58 |
| Nonprofit Organization Management | 76 | 0.59 | 10792 | 83.16 |
| Entertainment | 74 | 0.57 | 10866 | 83.73 |
| Electrical/Electronic Manufacturing | 73 | 0.56 | 10939 | 84.30 |
| Food & Beverages | 72 | 0.55 | 11011 | 84.85 |
| Cosmetics | 65 | 0.50 | 11076 | 85.35 |
| Airlines/Aviation | 63 | 0.49 | 11139 | 85.84 |
| Consumer Goods | 63 | 0.49 | 11202 | 86.32 |
| Consumer Electronics | 62 | 0.48 | 11264 | 86.80 |
| Medical Practice | 60 | 0.46 | 11324 | 87.26 |
| Public Relations and Communications | 58 | 0.45 | 11382 | 87.71 |
| Civic & Social Organization | 55 | 0.42 | 11437 | 88.13 |
| Market Research | 54 | 0.42 | 11491 | 88.55 |
| Transportation/Trucking/Railroad | 53 | 0.41 | 11544 | 88.96 |
| Restaurants | 52 | 0.40 | 11596 | 89.36 |
| Warehousing | 51 | 0.39 | 11647 | 89.75 |
| Broadcast Media | 50 | 0.39 | 11697 | 90.14 |
| Events Services | 50 | 0.39 | 11747 | 90.52 |
| Computer & Network Security | 49 | 0.38 | 11796 | 90.90 |
| Environmental Services | 49 | 0.38 | 11845 | 91.28 |
| Media Production | 48 | 0.37 | 11893 | 91.65 |
| Computer Networking | 44 | 0.34 | 11937 | 91.99 |
| Food Production | 44 | 0.34 | 11981 | 92.32 |
| Gambling & Casinos | 42 | 0.32 | 12023 | 92.65 |
| Pharmaceuticals | 42 | 0.32 | 12065 | 92.97 |
| Publishing | 39 | 0.30 | 12104 | 93.27 |
| Biotechnology | 38 | 0.29 | 12142 | 93.57 |
| Mechanical or Industrial Engineering | 37 | 0.29 | 12179 | 93.85 |
| Computer Hardware | 35 | 0.27 | 12214 | 94.12 |
| Utilities | 33 | 0.25 | 12247 | 94.37 |
| Graphic Design | 32 | 0.25 | 12279 | 94.62 |
| Printing | 30 | 0.23 | 12309 | 94.85 |
| Security and Investigations | 30 | 0.23 | 12339 | 95.08 |
| Research | 29 | 0.22 | 12368 | 95.31 |
| Venture Capital & Private Equity | 29 | 0.22 | 12397 | 95.53 |
| Information Services | 28 | 0.22 | 12425 | 95.75 |
| Aviation & Aerospace | 24 | 0.18 | 12449 | 95.93 |
| Farming | 24 | 0.18 | 12473 | 96.12 |
| Mental Health Care | 23 | 0.18 | 12496 | 96.29 |
| Sports | 23 | 0.18 | 12519 | 96.47 |
| Chemicals | 22 | 0.17 | 12541 | 96.64 |
| Government Administration | 22 | 0.17 | 12563 | 96.81 |
| Law Practice | 19 | 0.15 | 12582 | 96.96 |
| Medical Devices | 19 | 0.15 | 12601 | 97.10 |
| Outsourcing/Offshoring | 19 | 0.15 | 12620 | 97.25 |
| Writing and Editing | 19 | 0.15 | 12639 | 97.40 |
| Business Supplies and Equipment | 18 | 0.14 | 12657 | 97.53 |
| Fund-Raising | 16 | 0.12 | 12673 | 97.66 |
| Professional Training & Coaching | 14 | 0.11 | 12687 | 97.77 |
| Government Relations | 11 | 0.08 | 12698 | 97.85 |
| Higher Education | 11 | 0.08 | 12709 | 97.93 |
| Machinery | 11 | 0.08 | 12720 | 98.02 |
| Semiconductors | 11 | 0.08 | 12731 | 98.10 |
| Wholesale | 11 | 0.08 | 12742 | 98.19 |
| Architecture & Planning | 10 | 0.08 | 12752 | 98.27 |
| Law Enforcement | 10 | 0.08 | 12762 | 98.34 |
| Music | 10 | 0.08 | 12772 | 98.42 |
| Translation and Localization | 10 | 0.08 | 12782 | 98.50 |
| Civil Engineering | 9 | 0.07 | 12791 | 98.57 |
| Defense & Space | 9 | 0.07 | 12800 | 98.64 |
| Individual & Family Services | 9 | 0.07 | 12809 | 98.71 |
| Program Development | 9 | 0.07 | 12818 | 98.77 |
| Renewables & Environment | 9 | 0.07 | 12827 | 98.84 |
| Executive Office | 8 | 0.06 | 12835 | 98.91 |
| International Trade and Development | 8 | 0.06 | 12843 | 98.97 |
| Veterinary | 8 | 0.06 | 12851 | 99.03 |
| Industrial Automation | 7 | 0.05 | 12858 | 99.08 |
| Photography | 7 | 0.05 | 12865 | 99.14 |
| Public Safety | 7 | 0.05 | 12872 | 99.19 |
| Investment Management | 6 | 0.05 | 12878 | 99.24 |
| Motion Pictures and Film | 6 | 0.05 | 12884 | 99.28 |
| Primary/Secondary Education | 6 | 0.05 | 12890 | 99.33 |
| Religious Institutions | 6 | 0.05 | 12896 | 99.38 |
| Animation | 5 | 0.04 | 12901 | 99.41 |
| Capital Markets | 5 | 0.04 | 12906 | 99.45 |
| Import and Export | 5 | 0.04 | 12911 | 99.49 |
| Package/Freight Delivery | 5 | 0.04 | 12916 | 99.53 |
| Packaging and Containers | 5 | 0.04 | 12921 | 99.57 |
| Commercial Real Estate | 4 | 0.03 | 12925 | 99.60 |
| Fishery | 4 | 0.03 | 12929 | 99.63 |
| Investment Banking | 4 | 0.03 | 12933 | 99.66 |
| Luxury Goods & Jewelry | 4 | 0.03 | 12937 | 99.69 |
| Philanthropy | 4 | 0.03 | 12941 | 99.72 |
| Wireless | 4 | 0.03 | 12945 | 99.75 |
| Furniture | 3 | 0.02 | 12948 | 99.78 |
| Maritime | 3 | 0.02 | 12951 | 99.80 |
| Mining & Metals | 3 | 0.02 | 12954 | 99.82 |
| Performing Arts | 3 | 0.02 | 12957 | 99.85 |
| Plastics | 3 | 0.02 | 12960 | 99.87 |
| Public Policy | 3 | 0.02 | 12963 | 99.89 |
| Libraries | 2 | 0.02 | 12965 | 99.91 |
| Military | 2 | 0.02 | 12967 | 99.92 |
| Nanotechnology | 2 | 0.02 | 12969 | 99.94 |
| Textiles | 2 | 0.02 | 12971 | 99.95 |
| Alternative Dispute Resolution | 1 | 0.01 | 12972 | 99.96 |
| Museums and Institutions | 1 | 0.01 | 12973 | 99.97 |
| Ranching | 1 | 0.01 | 12974 | 99.98 |
| Shipbuilding | 1 | 0.01 | 12975 | 99.98 |
| Sporting Goods | 1 | 0.01 | 12976 | 99.99 |
| Wine and Spirits | 1 | 0.01 | 12977 | 100.00 |
| **Frequency Missing:** 4903 | | | | |

**Table 7:** PROC FREQ results for *function*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| Information Technology | 1749 | 15.31 | 1749 | 15.31 |
| Sales | 1468 | 12.85 | 3217 | 28.16 |
| Engineering | 1348 | 11.80 | 4565 | 39.96 |
| Customer Service | 1229 | 10.76 | 5794 | 50.71 |
| Marketing | 830 | 7.26 | 6624 | 57.98 |
| Administrative | 630 | 5.51 | 7254 | 63.49 |
| Design | 340 | 2.98 | 7594 | 66.47 |
| Health Care Provider | 338 | 2.96 | 7932 | 69.43 |
| Education | 325 | 2.84 | 8257 | 72.27 |
| Other | 325 | 2.84 | 8582 | 75.12 |
| Management | 317 | 2.77 | 8899 | 77.89 |
| Business Development | 228 | 2.00 | 9127 | 79.89 |
| Accounting/Auditing | 212 | 1.86 | 9339 | 81.74 |
| Human Resources | 205 | 1.79 | 9544 | 83.54 |
| Project Management | 183 | 1.60 | 9727 | 85.14 |
| Finance | 172 | 1.51 | 9899 | 86.64 |
| Consulting | 144 | 1.26 | 10043 | 87.90 |
| Art/Creative | 132 | 1.16 | 10175 | 89.06 |
| Writing/Editing | 132 | 1.16 | 10307 | 90.21 |
| Production | 116 | 1.02 | 10423 | 91.23 |
| Product Management | 114 | 1.00 | 10537 | 92.23 |
| Quality Assurance | 111 | 0.97 | 10648 | 93.20 |
| Advertising | 90 | 0.79 | 10738 | 93.99 |
| Business Analyst | 84 | 0.74 | 10822 | 94.72 |
| Data Analyst | 82 | 0.72 | 10904 | 95.44 |
| Public Relations | 76 | 0.67 | 10980 | 96.11 |
| Manufacturing | 74 | 0.65 | 11054 | 96.75 |
| General Business | 68 | 0.60 | 11122 | 97.35 |
| Research | 50 | 0.44 | 11172 | 97.79 |
| Legal | 47 | 0.41 | 11219 | 98.20 |
| Strategy/Planning | 46 | 0.40 | 11265 | 98.60 |
| Training | 38 | 0.33 | 11303 | 98.93 |
| Supply Chain | 36 | 0.32 | 11339 | 99.25 |
| Financial Analyst | 33 | 0.29 | 11372 | 99.54 |
| Distribution | 24 | 0.21 | 11396 | 99.75 |
| Purchasing | 15 | 0.13 | 11411 | 99.88 |
| Science | 14 | 0.12 | 11425 | 100.00 |
| **Frequency Missing:** 6455 | | | | |

From Tables 3-7, we can see that *employment\_type* has five categories with the most frequent category being “Full-time”, *required\_experience* has seven categories with the most frequent category being “Mid-Senior level”, and *required\_education* has 13 categories with the most frequent category being “Bachelor’s Degree.” However, some of the categories of *required\_education* are redundant. In particular, there are multiple categories of the “Vocational” type; it would be logical to merge these categories. As for the categorical variables *industry* and *function*, *industry* has 131 categories with the most frequent category being “Information Technology and Services” while *function* has 37 categories with the most frequent category being “Information Technology.”

**Cleaning the Data**

It is clear that several variables in the EMSCAD need to be cleaned. In particular, the following changes will be made:

* The character variable *location* stating the country, province and city of the job is too specific. Instead, only the country portion will be kept as a new categorical variable called *country*.
* The character variable *salary\_range* has too many missing values (83.96% according to Table 2) for the actual salary ranges to be useful. Instead, it will be 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 the categorical variable *required\_education* will be merged into a single category. Also, the categories “Some High School Coursework” and “High School or equivalent” will be merged since these two categories are similar.
* Missing values for the categorical variables *employment\_type*, *required\_experience*, *required\_education*, *industry,* *function*, and *country* will be assigned the category “Unspecified”.
* The character variable *department* will be dropped because it is similar to the categorical variable *function* but has a higher sparsity (*department* is 64.61% sparse compared to *function* which is only 36.10% sparse according to Table 2).
* The character variable *title* and the text variables *company\_profile*, *description*, *requirements,* and *benefits* will be dropped from the dataset. These variables are dealt with separately in the Text Mining section.

Here is the SAS code that was used to make these changes. The results were saved as a new SAS table called “EMSCAD\_clean”.

/\* Clean data \*/

**data** capstone.EMSCAD\_clean;

retain job\_id country employment\_type required\_experience required\_education industry function mentions\_salary telecommuting has\_company\_logo has\_questions fraudulent;

length employment\_type $15.;

format employment\_type $15.;

set capstone.EMSCAD;

/\* Extract country from location \*/

country=scan(location, **1**);

/\* Salary indicator \*/

if missing(salary\_range) then mentions\_salary=**0**;

else mentions\_salary=**1**;

/\* Clean categories of required\_education \*/

if required\_education =: 'Vocational' then required\_education='Vocational';

else if required\_education =: 'Some High School' then required\_education='High School or equivalent';

/\* Address missing values \*/

array x{\*} country &cat\_vars;

do \_n\_=**1** to dim(x);

if missing(x{\_n\_}) then x{\_n\_}='Unspecified';

end;

/\* Drop unnecessary columns \*/

drop title location department salary\_range company\_profile description requirements benefits;

**run**;

Another modification which will improve the appearance of certain graphs later on is reordering the levels of the categorical variables so that they appear in a more natural order. In particular, the categories of *employment\_type* can be ordered from highest commitment (“Full-time”) to least commitment (“Temporary”), the categories of *required\_experience* can be ordered from least experience (“Internship”) to most experience (“Executive”), and the categories of *required\_education* can be ordered from least education (“High School or equivalent”) to most education (“Doctorate”). Here is the SAS code to reorder the categorical variable levels:

/\* Create formats for reordering categorical variable levels \*/

**proc** **format**;

value etf **1**='Full-time'

**2**='Part-time'

**3**='Contract'

**4**='Temporary'

**5**='Other'

**6**='Unspecified';

value rexf **1**='Internship'

**2**='Entry level'

**3**='Associate'

**4**='Mid-Senior level'

**5**='Director'

**6**='Executive'

**7**='Not Applicable'

**8**='Unspecified';

value redf **1**='High School or equivalent'

**2**='Some college coursework'

**3**='Associate Degree'

**4**="Bachelor's Degree"

**5**="Master's Degree"

**6**='Doctorate'

**7**='Professional'

**8**='Vocational'

**9**='Certification'

**10**='Unspecified';

**run**;

/\* Reorder categorical variable levels \*/

**data** capstone.EMSCAD\_clean;

retain job\_id country employment\_type required\_experience required\_education industry function mentions\_salary telecommuting has\_company\_logo has\_questions fraudulent;

format employment\_type etf. required\_experience rexf.

required\_education redf.;

set capstone.EMSCAD\_clean(rename=(employment\_type=et

required\_experience=rex

required\_education=red));

select (et);

when ('Full-time') employment\_type=**1**;

when ('Part-time') employment\_type=**2**;

when ('Contract') employment\_type=**3**;

when ('Temporary') employment\_type=**4**;

when ('Other') employment\_type=**5**;

when ('Unspecified') employment\_type=**6**;

end;

select (rex);

when ('Internship') required\_experience=**1**;

when ('Entry level') required\_experience=**2**;

when ('Associate') required\_experience=**3**;

when ('Mid-Senior level') required\_experience=**4**;

when ('Director') required\_experience=**5**;

when ('Executive') required\_experience=**6**;

when ('Not Applicable') required\_experience=**7**;

when ('Unspecified') required\_experience=**8**;

end;

select (red);

when ('High School or equivalent') required\_education=**1**;

when ('Some college coursework') required\_education=**2**;

when ('Associate Degree') required\_education=**3**;

when ("Bachelor's Degree") required\_education=**4**;

when ("Master's Degree") required\_education=**5**;

when ('Doctorate') required\_education=**6**;

when ('Professional') required\_education=**7**;

when ('Vocational') required\_education=**8**;

when ('Certification') required\_education=**9**;

when ('Unspecified') required\_education=**10**;

end;

drop et rex red;

**run**;

**Collapsing Categorical Variable Levels**

The data is now clean, but there are still some potential problems which may arise with the categorical variables. As mentioned earlier, the standard way to incorporate a categorical variable into a logistic regression model is to encode the categories with numeric dummy variables. For categorical variables with a low number of categories such as *employment\_type*, *required\_experience*, and *required\_education*, this process should work fine, but for categorical variables with a high number of categories such as *industry,* *function*, and *country*, creating dummy variables to represent all of the categories is not practical and can lead to problems like high-dimensionality and quasi-complete separation (Patetta, Lesson 3.2). Therefore, it is necessary to collapse the levels of *industry*, *function*, and *country* before these variables can be used in the logistic regression model.

First, cross tabulations of *industry*, *function*, and *country* with *fraudulent* were performed to determine which categories have fraudulent job advertisements and which categories do not:

/\* Cross tabulations \*/

**proc** **freq** data=capstone.EMSCAD\_clean order=freq;

tables (industry function country)\*fraudulent;

**run**;

The results are quite lengthy and are not presented here for brevity’s sake (the interested reader may consult the Appendix). The cross tabulation of *country* with *fraudulent* shows that a large majority of fraudulent job advertisements (84.30%) come from the United States. Therefore, the levels of the categorical variable *country* were collapsed by replacing *country* with a binary variable *from\_US* which is true if the job advertisement comes from the United States or false if the job advertisement comes from any other country including “Unspecified”. Here is the SAS code that was used to create the *from\_US* variable:

/\* United States indicator \*/

**data** capstone.EMSCAD\_clean;

set capstone.EMSCAD\_clean;

if country='US' then from\_US=**1**;

else from\_US=**0**;

**run**;

On the other hand, *industry* and *function* do not possess a single category that dominates the others in terms of how many fraudulent job advertisements exist for that category. Therefore, the levels of the categorical variables *industry* and *function* were collapsed using the smooth weight-of-evidence (SWOE) technique discussed in (Patetta, Lesson 3.2). The SWOE technique converts *industry* and *function* into continuous numeric variables by replacing each category with the smoothed log-odds of the target event for that category, calculated using the formula

where is the number of fraudulent job advertisements for the category, is the number of legitimate job advertisements for the category, is the proportion of fraudulent job advertisements in the dataset (approximately 4.84%), and is a smoothing parameter selected by the analyst, chosen to be for the study.[[3]](#footnote-3) Here is the SAS code to perform the SWOE technique on *industry*, storing the results as a new variable called *industry\_SWOE*:

/\* Determine population proportion of fraudulent job ads \*/

%global rho1;

**proc** **sql**;

select mean(fraudulent) into: rho1

from capstone.EMSCAD;

**quit**;

/\* Count fraudulent job ads per industry \*/

**proc** **means** data=capstone.EMSCAD\_clean sum nway;

class industry;

var fraudulent;

output out=work.industry\_counts sum=events;

**run**;

/\* Compute industry SWOE \*/

**data** work.industry\_counts;

set work.industry\_counts;

industry\_SWOE = log((events + &rho1)/(\_FREQ\_ - events + (**1** - &rho1)));

**run**;

The same type of code was used to perform the SWOE technique on *function* and store the results as a new variable called *function\_SWOE*. The variables *industry\_SWOE* and *function\_SWOE* were then added to the “EMSCAD\_clean” dataset and the results were saved as a new SAS table called “EMSCAD\_clean\_SWOE”:

/\* Add industry\_SWOE and function\_SWOE to dataset \*/

**proc** **sql**;

create table capstone.EMSCAD\_clean\_SWOE as

select E.\*, I.industry\_SWOE, F.function\_SWOE

from capstone.EMSCAD\_clean E

left join work.industry\_counts I

on E.industry = I.industry

left join work.function\_counts F

on E.function = F.function

order by E.job\_id;

**quit**;

**Text Mining**

In this section, we switch to using Python to extract information from the character variable *title* and the text variables *company\_profile*, *description*, *requirements*, and *benefits*. First, all necessary functions were imported from the standard data science libraries NumPy, Pandas, Re and Matplotlib as well as the NLTK text mining library:

*# Import required libraries*

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **re**

**import** **matplotlib.pyplot** **as** **plt**

**import** **nltk**

**from** **nltk.corpus** **import** stopwords

**from** **nltk.tokenize** **import** word\_tokenize

**from** **nltk.stem** **import** WordNetLemmatizer

**from** **nltk.probability** **import** FreqDist

Next, the EMSCAD dataset was read into a Pandas dataframe called “df”, keeping only the variables needed for text mining and replacing any occurrences of NaN with blank strings.

*# Import data*

df = pd.read\_csv(path + '/EMSCAD.csv', engine='python')

df.drop(['location', 'department', 'salary\_range', 'telecommuting', 'has\_company\_logo', 'has\_questions', 'employment\_type', 'required\_experience', 'required\_education', 'industry', 'function'], inplace=**True**, axis=1)

*# Replace NaNs with blanks*

df.fillna(' ', inplace = **True**)

For each job advertisement, the text features *company\_profile*, *description*, *requirements,* and *benefits* were tokenized (transformed into a list of single-word units called tokens). The tokenizer function used in the study is based on the example given in (Sivarajah, 2020) and strips the text features of any punctuation or unnecessary words such as articles or conjunctions (collectively known as “stopwords” in the NLTK library) prior to tokenization taking place. The tokenizer also lemmatizes each token (puts the token into singular, present-tense form) using the built-in WordNetLemmatizer function from the NLTK library.

*# Create stopwords list*

stops = stopwords.words('english')

stops.remove('no')

stops += ['amp', 'aker']

*# Create tokenizer function*

lemmatizer = WordNetLemmatizer()

**def** tokenize(text):

*# Remove non-alphanumeric characters and normalize case*

text = re.sub(r'[^a-zA-Z0-9]', ' ', text.lower())

*# Tokenize text*

tokens = word\_tokenize(text)

*# Lemmatize tokens and remove stop words*

tokens = [lemmatizer.lemmatize(word) **for** word **in** tokens **if** word **not** **in** stops]

**return** tokens

*# Create tokens column*

**def** get\_tokens(x):

cpt = tokenize(x.company\_profile)

dt = tokenize(x.description)

rt = tokenize(x.requirements)

bt = tokenize(x.benefits)

**return** cpt + dt + rt + bt

df['tokens'] = df.apply(**lambda** x: [get\_tokens(x)],

axis=1, result\_type='expand')

Also, a column of bigrams (two-word phrases) was created from the “tokens” column:

*# Create bigrams column*

**def** get\_bigrams(token\_list):

bigram\_list = list()

**for** i **in** range(len(token\_list)-1):

bigram = token\_list[i] + ' ' + token\_list[i+1]

bigram\_list.append(bigram)

**return** bigram\_list

df['bigrams'] = df.apply(**lambda** x: [get\_bigrams(x.tokens)],

axis=1, result\_type='expand')

The tokens and bigrams for fraudulent and legitimate job advertisements were then sorted into separate lists for frequency analysis:

*# Create lists of tokens & bigrams from fraudulent/legitimate job ads*

fraud\_tokens, legit\_tokens = list(), list()

fraud\_bigrams, legit\_bigrams = list(), list()

**for** i **in** range(len(df)):

tok = df.tokens[i]

bi = df.bigrams[i]

**if** df.fraudulent[i] == 1:

fraud\_tokens += tok

fraud\_bigrams += bi

**else**:

legit\_tokens += tok

legit\_bigrams += bi

A frequency plot of the top 20 tokens in fraudulent job advertisements was generated using the NLTK FreqDist function discussed in Chapter 1 of (Bird et al., 2009):

*# Top 20 tokens in fraudulent job ads plot*

ft\_dist = FreqDist(fraud\_tokens).most\_common(20)

top\_ft, ft\_freq = list(), list()

**for** i **in** range(20):

top\_ft.append(ft\_dist[-(i+1)][0])

ft\_freq.append(ft\_dist[-(i+1)][1])

plt.barh(top\_ft, ft\_freq, color='red', alpha=.7)

plt.title('Top 20 Tokens in Fraudulent Job Advertisements')

plt.ylabel('Token')

plt.xlabel('Frequency')

**for** i **in** range(20):

plt.text(ft\_freq[i] + 5, i, ft\_freq[i], color='black',

ha='left', va='center')

plt.show()

Chart, bar chart

Description automatically generated**Figure 1:** Top 20 tokens in fraudulent job advertisements.

Similar code was used to generate frequency plots of the top 20 tokens in legitimate job advertisements as well as the top 20 bigrams in fraudulent and legitimate job advertisements. These plots are shown in Figures 2-4.

**Chart, bar chart

Description automatically generatedFigure 2:** Top 20 tokens in legitimate job advertisements.

**Figure 3:** Top 20 bigrams in fraudulent job advertisements.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated**Figure 4:** Top 20 bigrams in legitimate job advertisements.

There are some interesting observations to make about Figures 1-4. The fraudulent job advertisements in the EMSCAD do not appear to be significantly different from the legitimate job advertisements in terms of single-word tokens. In Figures 1 and 2, we can see that words such as “work”, “experience”, “skill”, “customer”, and “company” are frequently found in both types of advertisements.

However, the fraudulent job advertisements are noticeably different from the legitimate job advertisements when it comes to bigrams. According to Figure 3, fraudulent job advertisements often contain bigrams such as “data entry”, “no experience”, “work home”, “part time”, “signing bonus”, “get started”, and “high school” which do not appear in Figure 4, the frequency plot for the top 20 bigrams in legitimate job advertisements. These bigrams imply that many of the fraudulent job advertisements in the EMSCAD are work-from-home positions requiring little to no experience and offering signing bonuses to lure applicants. Conversely, the bigrams “fast paced”, “team member”, “5 year”, “long term”, “high quality”, “fast growing”, and “competitive salary” are unique to legitimate job advertisements. These bigrams indicate that legitimate job advertisements are looking for applicants with long-term commitment and collaboration skills required for positions at actual companies.

Since the presence of certain bigrams may be an effective way of differentiating between fraudulent and legitimate job advertisements, a new binary variable called *has\_fraud­\_bigram* was created which is true if the job advertisement contains a bigram from the “common\_fb” list shown below:

*# Fraudulent bigram indicator*

common\_fb = ['data entry', 'oil gas', 'gas industry', 'no experience', 'experience required', 'work home', 'part time', 'signing bonus', 'get started', 'high school', 'school diploma']

**def** has\_fb(bigram\_list):

**if** any([bigram **in** common\_fb **for** bigram **in** bigram\_list]):

**return** 1

**else**:

**return** 0

df['has\_fraud\_bigram'] = df.apply(**lambda** x: [has\_fb(x.bigrams)], axis=1, result\_type='expand')

Similarly, a new binary variable called *has\_legit\_bigram* was created which is true if the job advertisement contains a bigram from the “common\_lb” list shown below:

*# Legitimate bigram indicator*

common\_lb = ['full time', '5 year', '3 year', '2 year', 'long term', 'team member', 'high quality', 'fast growing', 'fast paced', 'ideal candidate', 'competitive salary']

**def** has\_lb(bigram\_list):

**if** any([bigram **in** common\_lb **for** bigram **in** bigram\_list]):

**return** 1

**else**:

**return** 0

df['has\_legit\_bigram'] = df.apply(**lambda** x: [has\_lb(x.bigrams)], axis=1, result\_type='expand')

In addition to tokens and bigrams, the presence of email addresses, phone numbers, or links to external websites in a job advertisement could be significant predictors of fraudulent status in the logistic regression model. Hence, the binary variables *has\_email*, *has\_phone*, and *has\_url* were created which are true if any of the job advertisement’s four text features (*company\_profile*, *description*, *requirements*, *benefits*) contain an email address, phone number, or URL, respectively. The detection of these attributes was made possible by the fact that the creators of the EMSCAD masked email addresses, phone numbers and URLs with the strings “#EMAIL”, “#PHONE”, and “#URL”, respectively.

*# Merge text features function*

**def** merge\_text(x):

cp = x.company\_profile

d = x.description

r = x.requirements

b = x.benefits

**return** cp + d + r + b

*# Email indicator*

**def** email(text):

**if** '#EMAIL' **in** text:

**return** 1

**else**:

**return** 0

df['has\_email'] = df.apply(**lambda** x: [email(merge\_text(x))],

axis=1, result\_type='expand')

*# Phone number indicator*

**def** phone(text):

**if** '#PHONE' **in** text:

**return** 1

**else**:

**return** 0

df['has\_phone'] = df.apply(**lambda** x: [phone(merge\_text(x))],

axis=1, result\_type='expand')

*# External URL indicator*

**def** url(text):

**if** '#URL' **in** text:

**return** 1

**else**:

**return** 0

df['has\_url'] = df.apply(**lambda** x: [url(merge\_text(x))],

axis=1, result\_type='expand')

Also, a binary variable called *money\_in\_title* was created which is true if the character variable *title* contains a “$” symbol. Past studies have shown that fraudulent job advertisements frequently mention salaries or wages in the title to lure applicants (Vidros et al., 2017), so the variable *money\_in\_title* has the potential to be a strong predictor of fraudulent status in the logistic regression model.

*# Money in title indicator*

**def** money\_in\_title(text):

**if** '$' **in** text:

**return** 1

**else**:

**return** 0

df['money\_in\_title'] = df.apply(**lambda** x: [money\_in\_title(x.title)],

axis=1, result\_type='expand')

Finally, for each job advertisement, the number of words in the text features *company\_profile*, *description*, *requirements*, and *benefits* were calculated and stored as the variables *company\_profile\_length*, *description\_length*, *requirements\_length*, and *benefits\_length*.

*# Word count function*

**def** word\_count(text):

*# Remove non-alphanumeric characters*

text = re.sub(r'[^a-zA-Z0-9]', ' ', text)

**return** len(text.split())

df['company\_profile\_length'] = df.apply(

**lambda** x: [word\_count(x.company\_profile)],

axis=1, result\_type='expand')

df['description\_length'] = df.apply(

**lambda** x: [word\_count(x.description)],

axis=1, result\_type='expand')

df['requirements\_length'] = df.apply(

**lambda** x: [word\_count(x.requirements)],

axis=1, result\_type='expand')

df['benefits\_length'] = df.apply(

**lambda** x: [word\_count(x.benefits)],

axis=1, result\_type='expand')

Once all of the new numeric variables had been created, any old variables that were no longer relevant were dropped from the dataset. The text mining results were saved as a CSV file so that they could be imported in SAS Studio.

*# Drop columns that are no longer needed*

df.drop(['title', 'company\_profile', 'description', 'requirements', 'benefits', 'tokens', 'bigrams', 'fraudulent'],

inplace=**True**, axis=1)

*# Export text mining results as CSV file*

df.to\_csv(path + '/text\_mining\_results.csv', index=**False**)

**List of New Variables**

Back in SAS Studio, the Python text mining results were imported and merged with the “EMSCAD\_clean\_SWOE” dataset using the following code:

/\* Import text mining results \*/

**proc** **import** datafile="&path/text\_mining\_results.csv"

dbms=csv

out=capstone.text\_mining\_results;

guessingrows=max;

**run**;

/\* Merge EMSCAD\_clean\_SWOE and text\_mining\_results \*/

%let final\_vars=employment\_type, required\_experience, required\_education, company\_profile\_length, description\_length, requirements\_length, benefits\_length, industry\_SWOE, function\_SWOE, from\_US, money\_in\_title, mentions\_salary, telecommuting, has\_company\_logo, has\_questions, has\_email, has\_phone, has\_url, has\_fraud\_bigram, has\_legit\_bigram, fraudulent;

**proc** **sql**;

create table capstone.EMSCAD\_final as

select &final\_vars

from capstone.EMSCAD\_clean\_SWOE E

join capstone.text\_mining\_results T

on E.job\_id=T.job\_id;

**quit**;

The resulting “EMSCAD\_final” dataset contains the following variables:

**Table 8:** 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 independent variables (*employment\_type*, *required\_experience*, *required\_education*, *company\_profile\_length*, *description\_length*, *requirements\_length*, *benefits­\_length*, *industry\_SWOE*, *function\_SWOE*, *from\_US*, *money\_in\_title*, *mentions\_salary*, *telecommuting*, *has\_company\_logo*, *has\_questions*, *has\_email*, *has\_phone*, *has\_url*, *has\_fraud\_bigram*, *has\_legit\_bigram*) have now been thoroughly prepared and can all be used directly to predict the value of *fraudulent*, the binary dependent variable for the logistic regression model.

ANALYSIS

**Visualizations**

In SAS Studio, the first type of analysis performed on the EMSCAD\_final dataset was creating several types of charts and graphs to visualize the distribution of fraudulent job advertisements across the independent variables. These charts and graphs provide many useful insights into the distinguishing features of fraudulent job advertisements. First, the following code was used to create a stacked bar chart showing the distribution of *employment\_type* categories grouped by fraudulent status along with a regular bar chart showing the fraudulent rates within each *employment\_type* category.

/\* Create Yes/No format for binary variables \*/

**proc** **format**;

value yn **1**='Yes' **0**='No';

**run**;

/\* employment\_type distribution \*/

title 'employment\_type Distribution';

**proc** **sgplot** data=capstone.EMSCAD\_final;

format fraudulent yn.;

vbar employment\_type / group=fraudulent groupdisplay=stack

datalabel seglabel seglabelattrs=(size=**4**) seglabelfitpolicy=noclip;

xaxis discreteorder=data;

yaxis label='Frequency' grid;

keylegend / title='Fraudulent' location=inside position=topright;

**run**;

/\* employment\_type fraudulent rates \*/

**proc** **freq** data=capstone.EMSCAD\_final noprint;

tables employment\_type\*fraudulent / out=FreqOut(where=(fraudulent=**1**)) outpct;

**run**;

**data** FreqOut;

set FreqOut;

row\_proportion=PCT\_ROW/**100**;

keep employment\_type row\_proportion;

**run**;

title 'employment\_type Fraudulent Rates';

**proc** **sgplot** data=FreqOut;

format row\_proportion percent6.2;

vbar employment\_type / response=row\_proportion datalabel

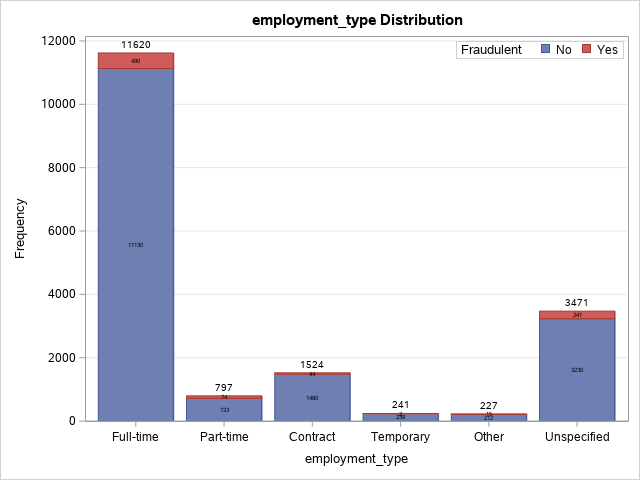
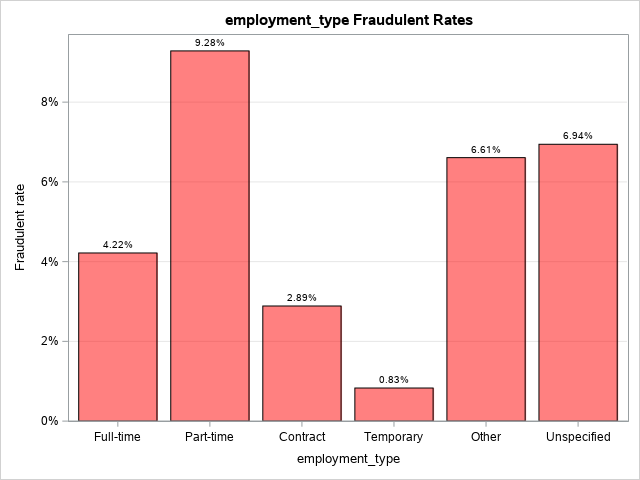
fillattrs=(color=red transparency=**0.5**);

xaxis discreteorder=data;

yaxis label='Fraudulent rate' grid;

**run**;

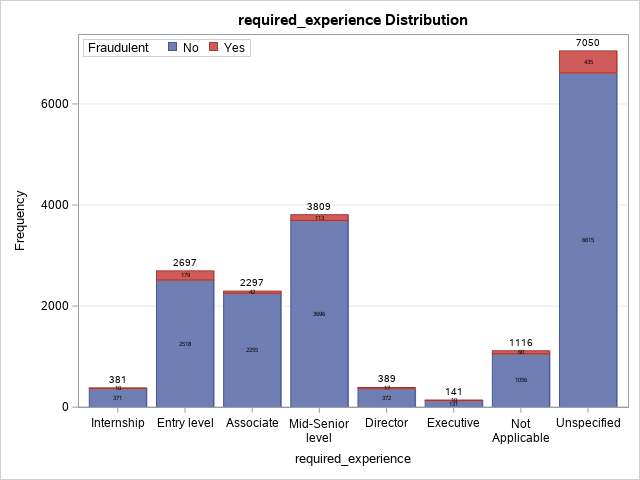
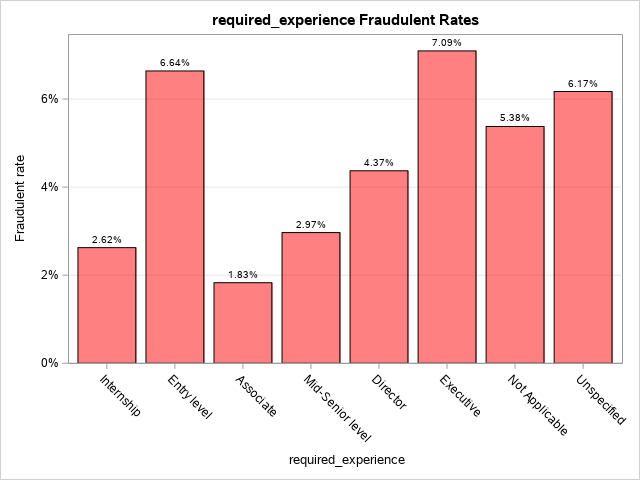
**Figure 5:** *employment\_type* distribution and fraudulent rates.



According to Figure 5, a majority of the job advertisements in the EMSCAD\_final dataset (11,620 advertisements representing 64.99% of the data) are full-time positions. Full-time positions also have 490 fraudulent cases which is more than any other employment type. However, looking at the fraudulent rates within each *employment\_type* category, “Part-time” has the highest fraudulent rate (9.28%) followed by “Unspecified” in second place (6.94%) and “Other” in third place (6.61%). Therefore, we can conclude that even though full-time positions are the most common and have the most fraudulent cases by sheer numbers, jobs where the commitment level is low (“Part-time”, “Other”, or “Unspecified”) are the riskiest in terms of scam likelihood.

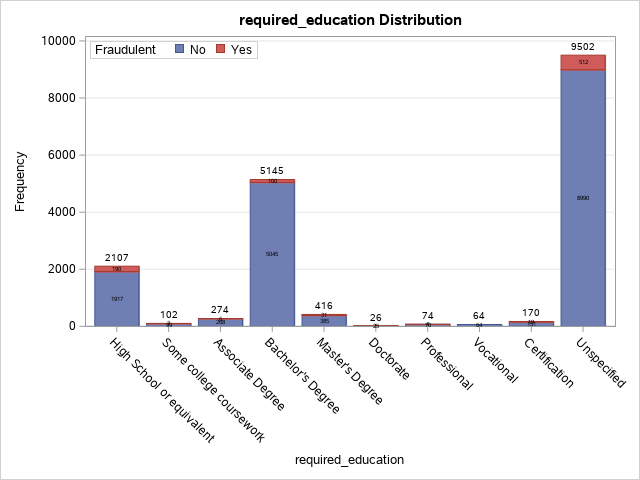
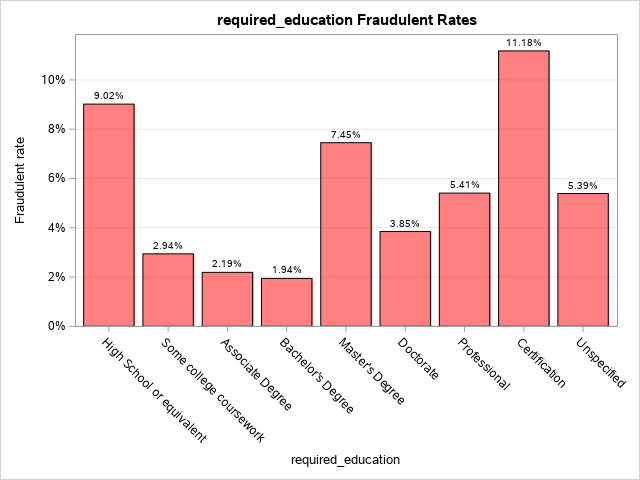
Similar code was used to generate distribution and fraudulent rate plots for the other two categorical variables, *required\_experience* and *required\_education* (Figures 6 & 7).

**Figure 6:** *required\_experience* distribution and fraudulent rates.



According to Figure 6, the “Unspecified” category where required experiencewas left blank occurs 7,050 times representing 39.43% of the data, making it the most common required experience level in the EMSCAD\_final dataset. The “Unspecified” category also has 435 fraudulent cases which is more than any other *required\_experience* category. As for fraudulent rates, the “Executive” category has the highest fraudulent rate (7.09%) followed by “Entry level” in second place (6.64%) and “Unspecified” in third place (6.17%). Thus, it appears that job advertisements where required experience is either very high (the “Executive” level) or very low (“Unspecified” or “Entry level”) are more likely to be scams than job advertisements with moderate required experience levels (“Associate” or “Mid-Senior level”).

**Figure 7:** *required\_education* distribution and fraudulent rates.



In Figure 7, the *required\_education* distribution shows that the “Unspecified” category where required education was left blank occurs 9,502 times representing a majority (53.14%) of the EMSCAD\_final dataset, followed by “Bachelor’s Degree” in second place (5,145 occurrences) and “High School or equivalent” in third place (2,107 occurrences). The remaining seven categories combined occur 1,126 times and only represent 6.30% of the data. The “Unspecified” category also has the most scams (512). However, looking at the *required\_education* fraudulent rates, the “Certification” category has the highest fraudulent rate (11.18%) followed by “High School or equivalent” (9.02%). These results are logical: a job advertisement scam is unlikely to ask for candidates with high education levels because that would deter certain groups of people from applying, and scammers want to reach as wide of an audience as possible in order to maximize their chances of success.

Next, the four text feature length variables *company\_profile\_length*, *description\_length*, *requirements\_length*, and *benefits\_length* were analyzed with histograms and box plots (Figures 8-11). The code shown below was used to create the *company\_profile\_length* histogram and box plot; the histograms and box plots for the other three variables were generated in a similar fashion. The SAS article (Matange, 2015) was referenced for the box plot code.

/\* company\_profile\_length histogram \*/

title 'company\_profile\_length Histogram';

**proc** **sgplot** data=capstone.EMSCAD\_final;

histogram company\_profile\_length / scale=count;

yaxis label='Frequency' grid;

**run**;

/\* company\_profile\_length box plot \*/

options validvarname=v7;

ods output sgplot=cpl\_boxplotdata(rename=(

BOX\_COMPANY\_PROFILE\_LENGTH\_X\_\_\_Y=value

BOX\_COMPANY\_PROFILE\_LENGTH\_X\_\_ST=stat

BOX\_COMPANY\_PROFILE\_LENGTH\_X\_\_\_X=cat));

**proc** **sgplot** data=capstone.EMSCAD\_final;

vbox company\_profile\_length / category=fraudulent;

**run**;

**data** cpl\_merged;

format fraudulent cat yn.;

set capstone.EMSCAD\_final

cpl\_boxplotdata(where=(value ne . and

stat in ('MIN' 'Q1' 'MEDIAN' 'Q3' 'MAX' 'MEAN')));

value=round(value, **0.01**);

**run**;

title 'company\_profile\_length Box Plot';

**proc** **sgplot** data=cpl\_merged noautolegend;

vbox company\_profile\_length / category=fraudulent group=fraudulent

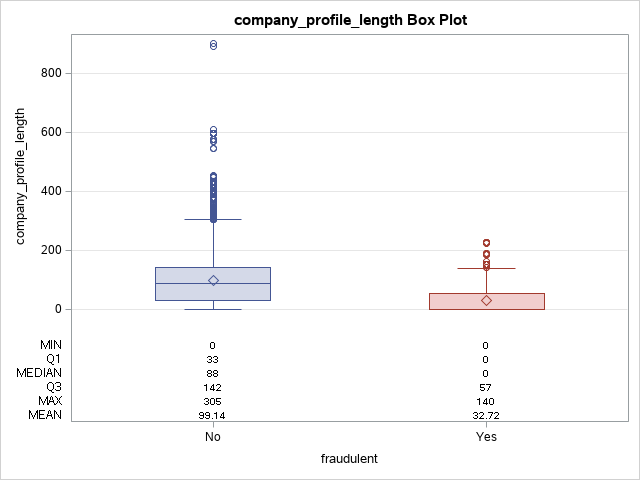
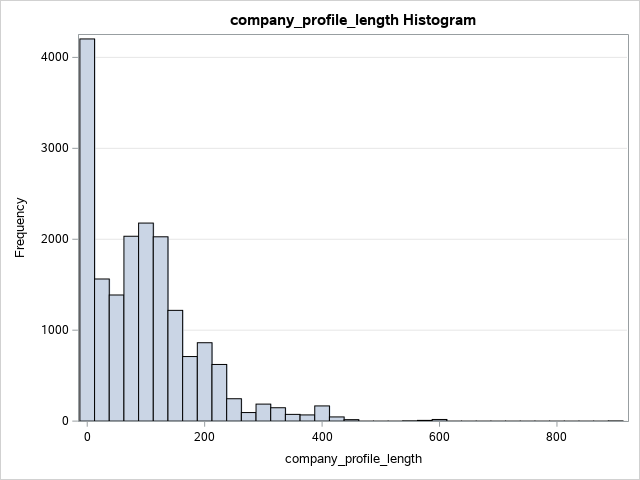
fillattrs=(transparency=**0.7**) meanattrs=(symbol=diamond);

xaxistable value / x=cat class=stat location=inside;

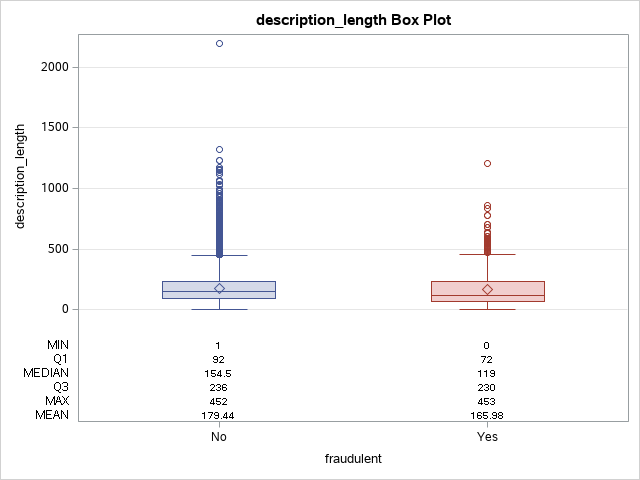
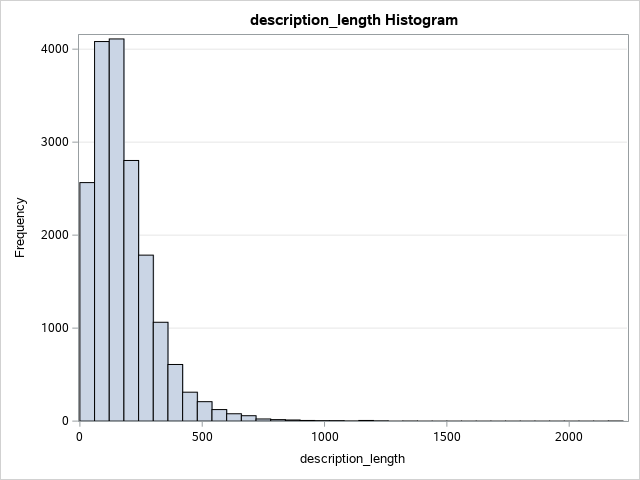
yaxis grid;

**run**;

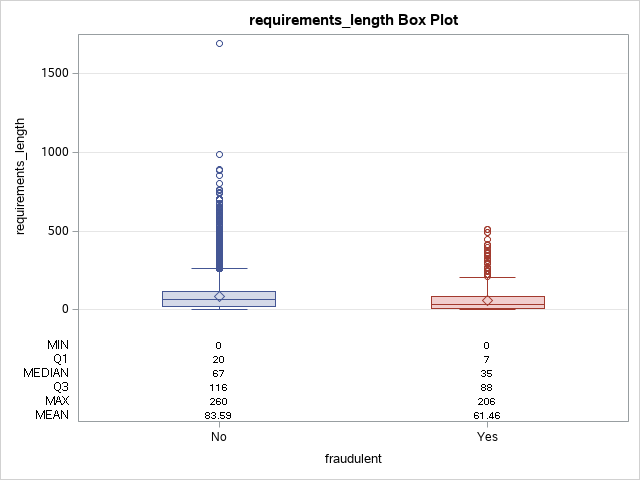
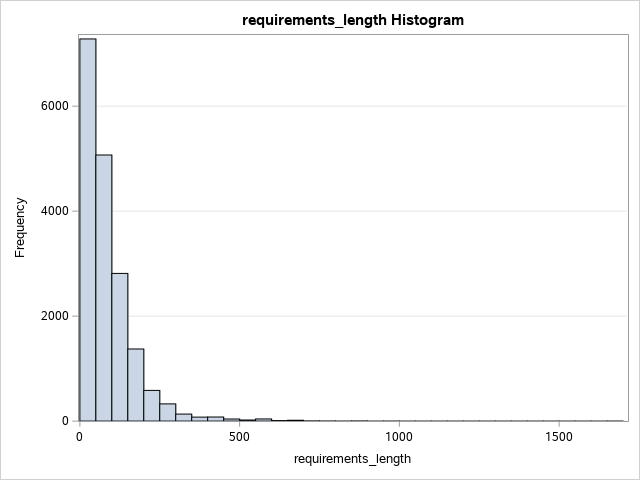
**Figure 8:** *company\_profile\_length* histogram and box plot.



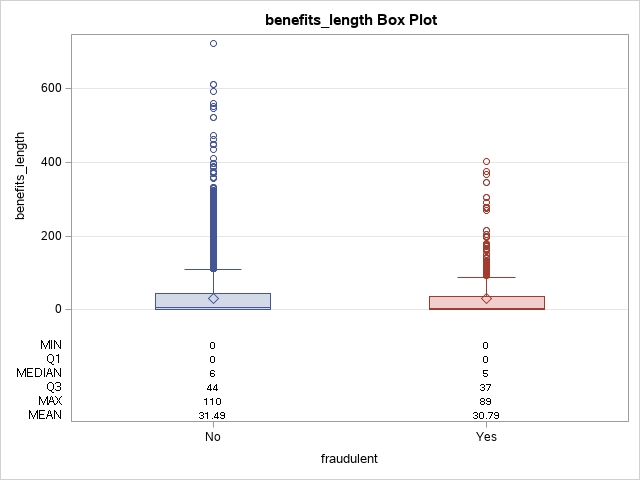
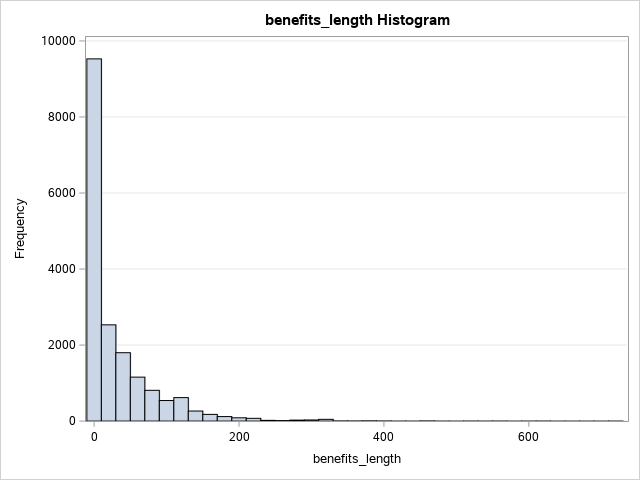
**Figure 9:** *description\_length* histogram and box plot.



**Figure 10:** *requirements\_length* histogram and box plot.



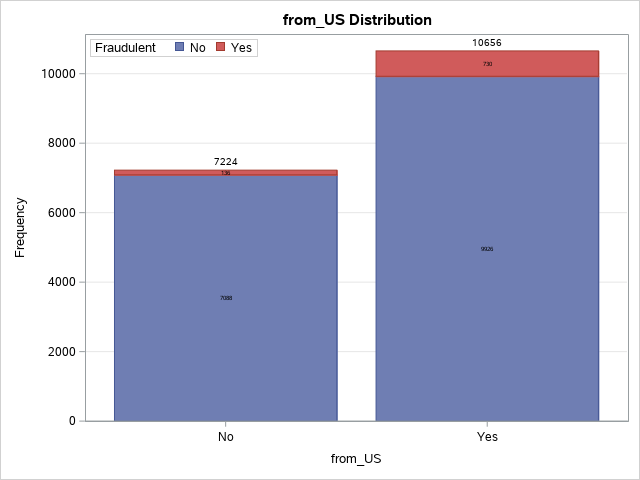
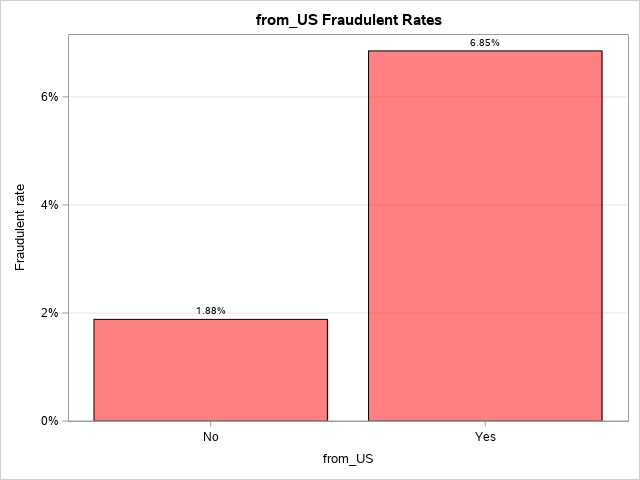
**Figure 11:** *benefits\_length* histogram and box plot.



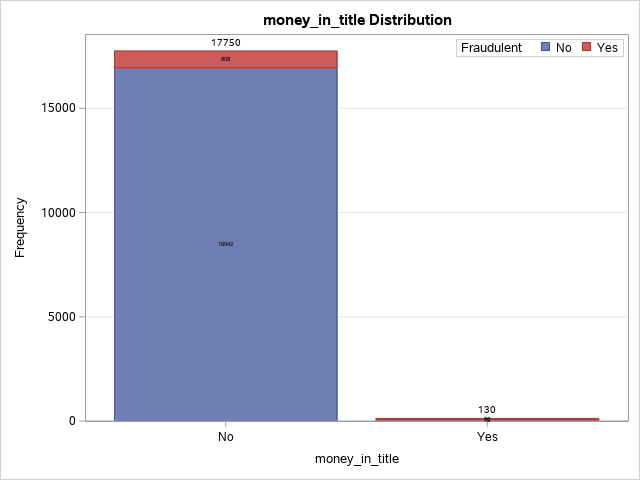
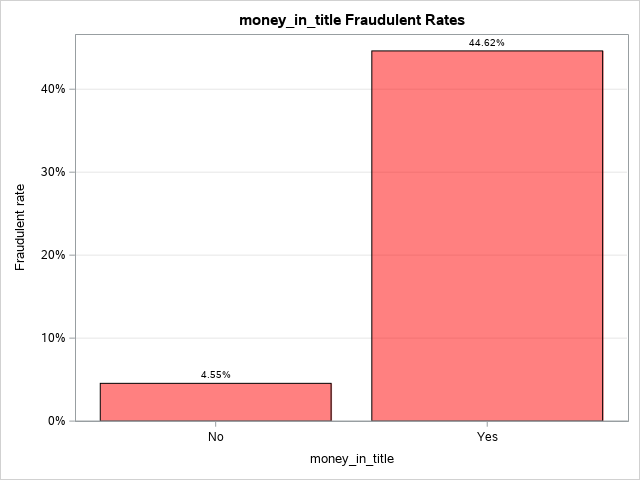
Looking at the histograms in Figures 8-11, the distributions of *company\_profile\_length*, *description\_length*, *requirements\_length* and *benefits\_length* are all skewed to the right which means that there are more job advertisements in the EMSCAD\_final dataset with short or blank text features than there are job advertisements with long, detailed text features. In addition, the box plots in Figures 8-11 indicate that fraudulent job advertisements are consistently shorter than legitimate job advertisements across all four text features. In particular, legitimate job advertisements have a mean company profile length of 99.14 words while fraudulent job advertisements have a mean company profile length of 32.72 words; legitimate job advertisements have a mean description length of 179.44 words while fraudulent job advertisements have a mean description length of 165.98 words; legitimate job advertisements have a mean requirements length of 83.59 words while fraudulent job advertisements have a mean requirements length of 61.46 words; and last but not least, legitimate job advertisements have a mean benefits length of 31.49 words while fraudulent job advertisements have a mean benefits length of 30.79 words. Clearly, shorter job advertisements are more likely to be scams than longer job advertisements.

Finally, the distributions and fraudulent rates of the binary variables in the EMSCAD\_final dataset were visualized with the same code that was used for the categorical variable plots in Figures 5-7. The binary variable plots are shown in Figures 12-22.

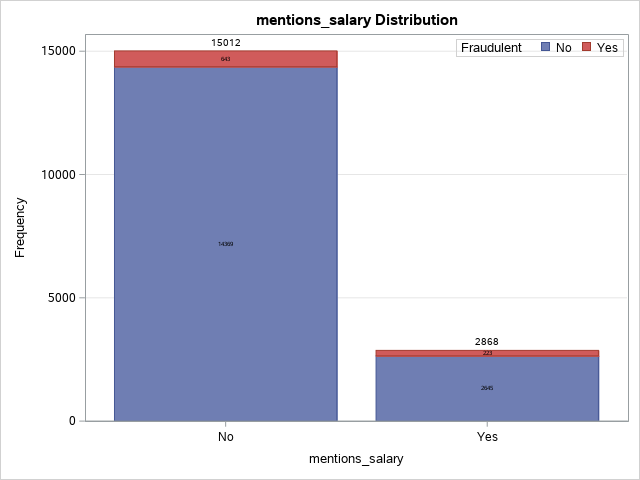
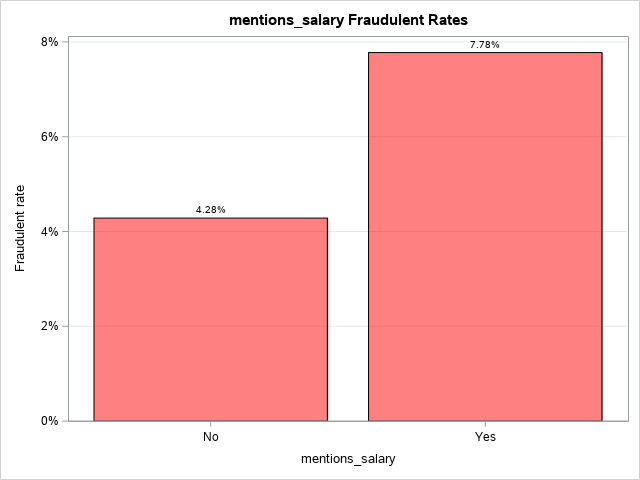
**Figure 12:** *from\_US* distribution and fraudulent rates.



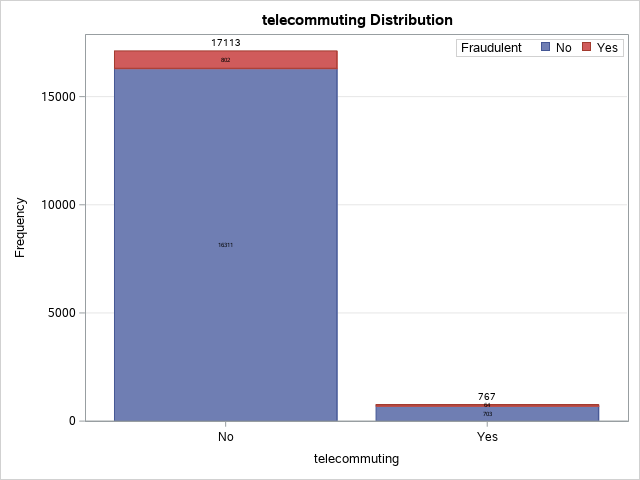
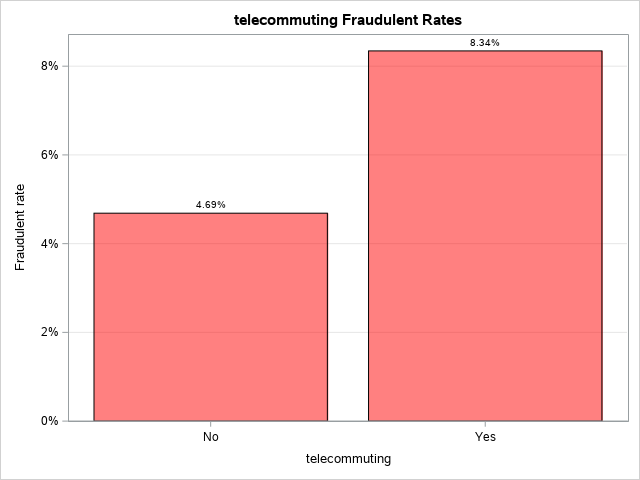
**Figure 13:** *money\_in\_title* distribution and fraudulent rates.



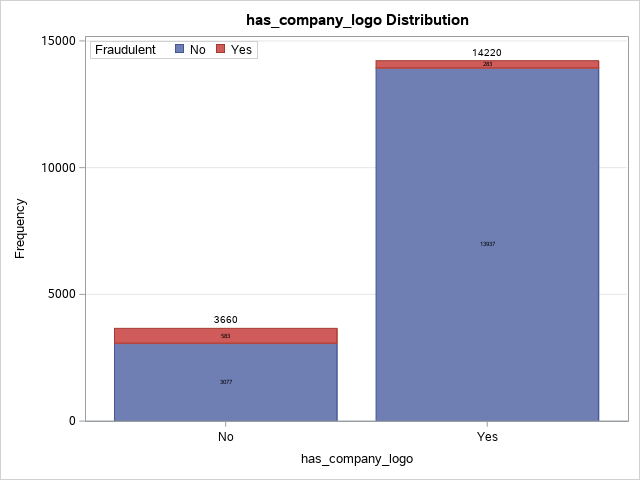
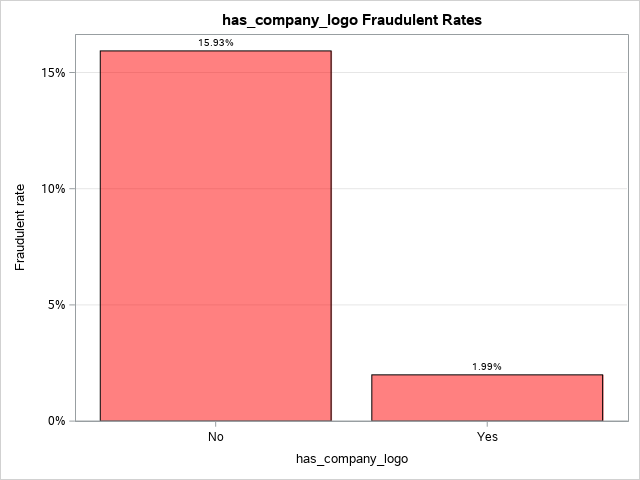
**Figure 14:** *mentions\_salary* distribution and fraudulent rates.



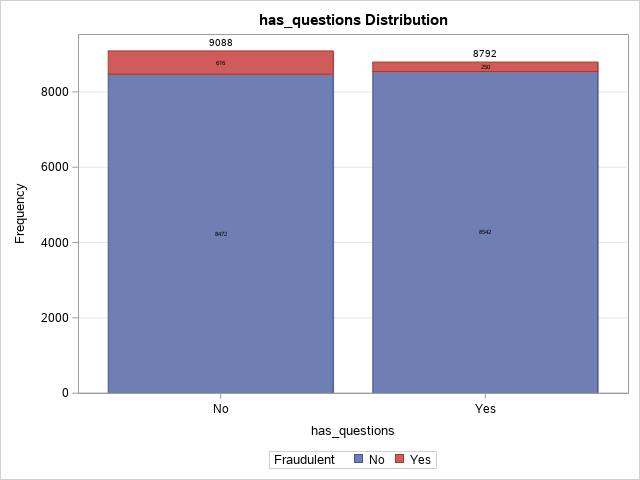
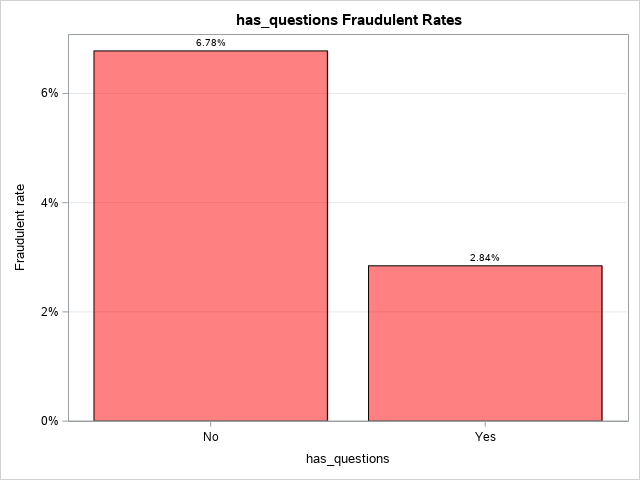
**Figure 15:** *telecommuting* distribution and fraudulent rates.



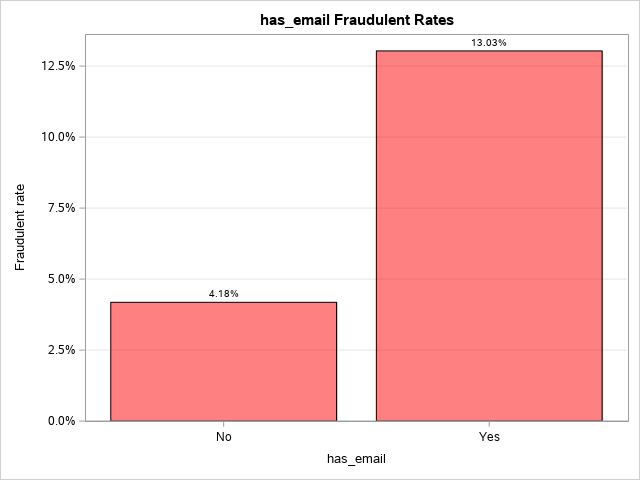
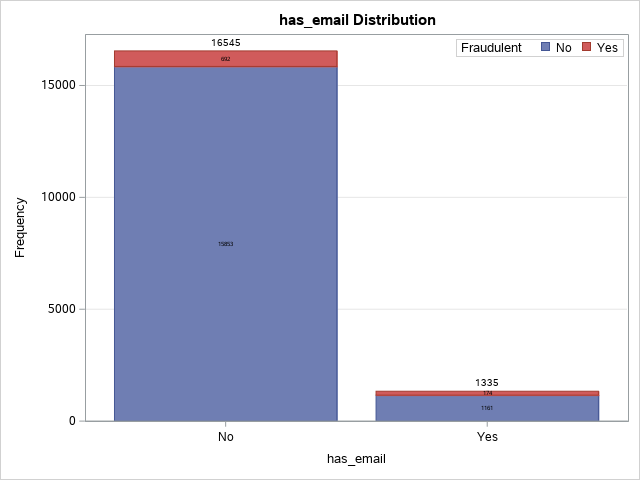
**Figure 16:** *has\_company\_logo* distribution and fraudulent rates.



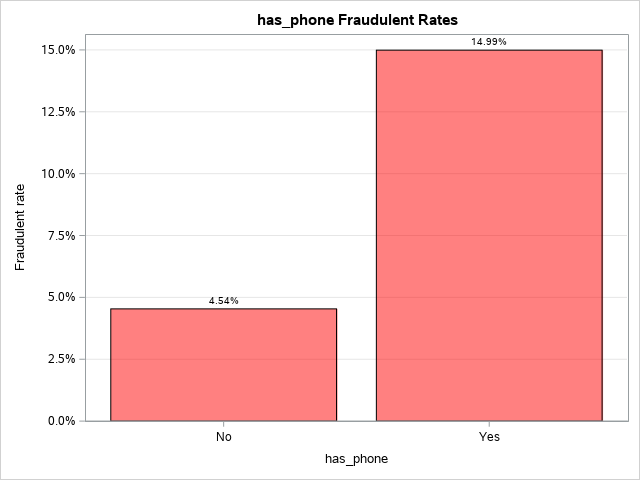
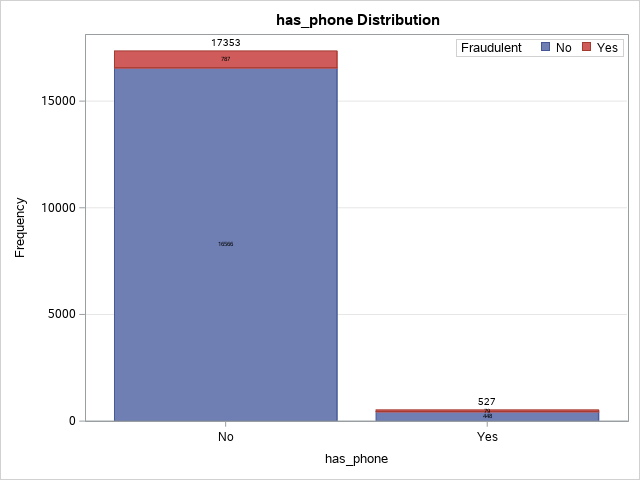
**Figure 17:** *has\_questions* distribution and fraudulent rates.



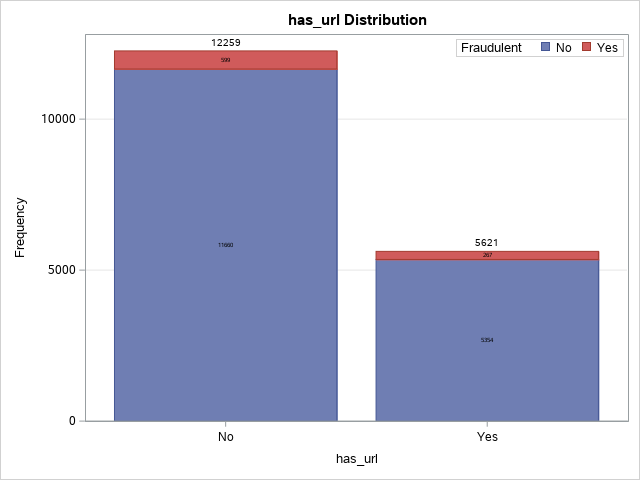
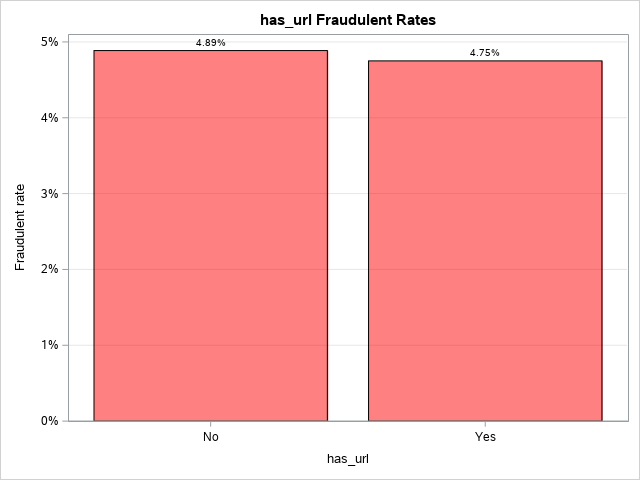
**Figure 18:** *has\_email* distribution and fraudulent rates.



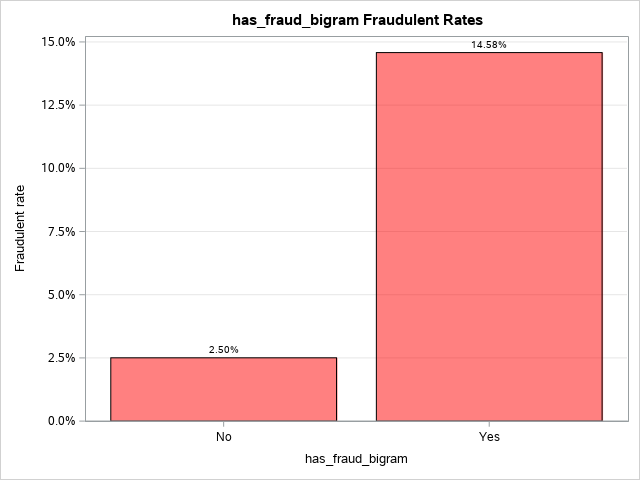
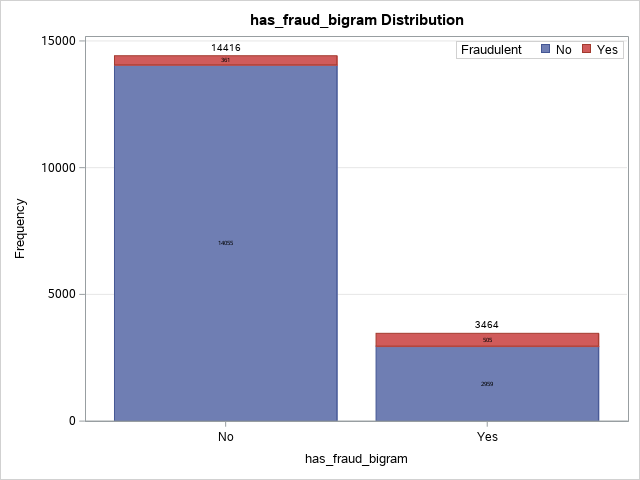
**Figure 19:** *has\_phone* distribution and fraudulent rates.



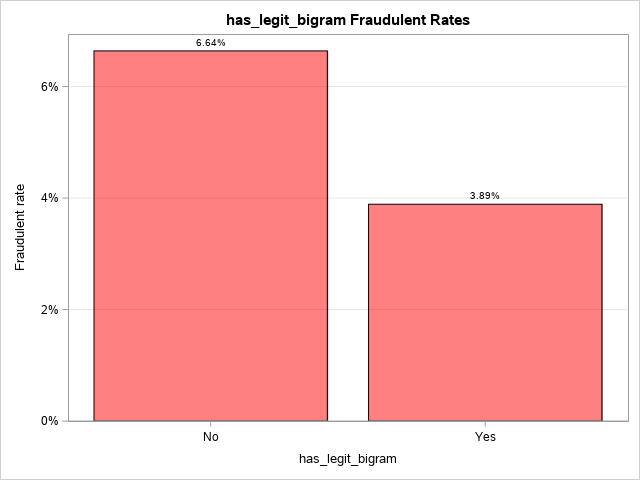
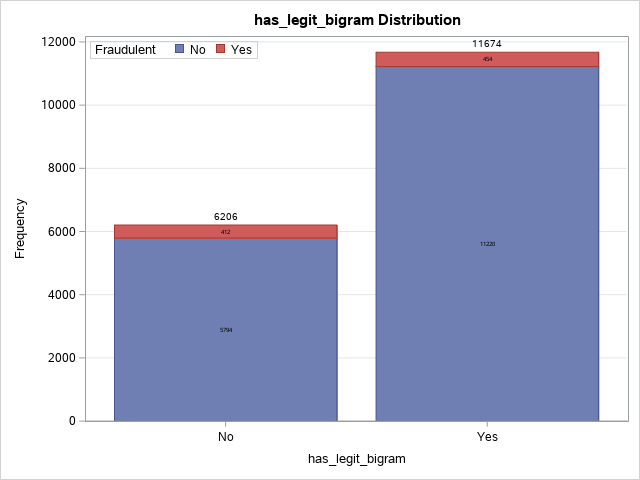
**Figure 20:** *has\_url* distribution and fraudulent rates.



**Figure 21:** *has\_fraud\_bigram* distribution and fraudulent rates.



**Figure 22:** *has\_legit\_bigram* distribution and fraudulent rates.



Here are several key observations about Figures 12-22:

* According to Figure 13, the vast majority of the job advertisements in the EMSCAD\_final dataset (17,750 advertisements representing 99.27% of the data) do not contain a “$” symbol in the title. However, the job advertisements which do have a dollar sign in the title have a fraudulent rate of 44.62% which is the highest fraudulent rate of any independent variable that has been examined so far. Hence, even though job advertisements which mention money in the title are rare, when they do occur, there is a high probability that the advertisement is a scam.
* Looking at Figures 12, 14, 15, 18, and 19, 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, from Figures 16 and 17, when a job advertisement has a company logo or screening questions, the probability that the advertisement is a scam decreases.
* In Figure 20, the presence of a link to an external website does not appear to have a significant effect on fraudulent status. The fraudulent rate of job advertisements which do not have URLs (4.89%) is roughly the same as the fraudulent rate of job advertisements which do have URLs (4.75%).
* According to Figure 21, job advertisements which contain a common fraud bigram have a higher fraudulent rate (14.58%) than job advertisements which do not (2.50%). This is a good sign that the *has\_fraud\_bigram* variable is meaningful. Likewise, in Figure 22, job advertisements which contain a common legitimate bigram have a lower fraudulent rate (3.89%) than job advertisements which do not (6.64%). Hence, *has\_legit\_bigram* may also be a good predictor of fraudulent status in the logistic regression model.

**Training and Validation Datasets**

A standard practice for creating a logistic regression model (and most predictive models for that matter) is to split the data into training and validation datasets. The training dataset is used to form various candidate models while the validation dataset is used for model comparison and selecting the best model. However, when the target event is rare in the initial dataset, it may be necessary to oversample observations with the target event prior to forming the training and validation datasets (Patetta, Lesson 1.2). This is certainly the case with the EMSCAD\_final dataset where fraudulent job advertisements only make up 4.84% of the data. The predicted probabilities of the logistic regression model will then need to be adjusted for oversampling later on. Here is the code that was used to oversample the EMSCAD\_final dataset:

/\* Sort EMSCAD\_final by fraudulent \*/

**proc** **sort** data=capstone.EMSCAD\_final;

by fraudulent;

**run**;

/\* Oversample fraudulent cases to form development dataset \*/

**proc** **surveyselect** data=capstone.EMSCAD\_final

method=srs n=(**2598**, **866**) seed=**6302021**

out=capstone.development(drop=SelectionProb SamplingWeight);

strata fraudulent;

**run**;

The development dataset contains all 866 fraudulent job advertisements and 2,598 randomly selected legitimate job advertisements from the EMSCAD\_final dataset so that fraudulent advertisements now make up 25% of the data, a percentage that should be sufficiently large for logistic regression (Bhalla, 2015). In order to make the results reproducible, the SEED option was enabled to fix the seed of the random number generator used by the SURVEYSELECT procedure to randomly select the legitimate job advertisements. From here, the training and validation datasets were formed with the following code:

/\* Take stratified sample of development dataset \*/

**proc** **surveyselect** data=capstone.development samprate=**0.7**

seed=**6302021** out=capstone.sample outall;

strata fraudulent;

**run**;

/\* Split sample into training & validation datasets \*/

**data** capstone.training(drop=Selected SelectionProb SamplingWeight)

capstone.validation(drop=Selected SelectionProb SamplingWeight);

set capstone.sample;

if selected then output capstone.training;

else output capstone.validation;

**run**;

In the second SURVEYSELECT procedure, the SAMPRATE=0.7 option randomly assigns 70% of the development dataset to the training dataset and the remaining 30% to the validation dataset. The STRATA statement ensures that the 25% proportion of fraudulent cases is maintained in both the training and validation datasets. Again, the SEED option was turned on to make the data partitioning process reproducible.

**All-Variables Model**

Now that the training and validation datasets have been formed, we are finally ready to create a logistic regression model to predict whether a job advertisement is fraudulent based on the various advertisement features in the EMSCAD\_final dataset (Table 8). To be clear, logistic regression is an appropriate modeling technique because other techniques such as linear regression require the dependent variable to be a continuous numeric variable, whereas *fraudulent* is binary (Patetta, Lesson 2.1). The first logistic regression model that was explored was an all-variables model where all 20 independent variables in the EMSCAD\_final dataset were included in the model. Here is the code that was used to create the all-variables model:

/\* List of all predictors \*/

%let all\_vars=employment\_type required\_experience required\_education company\_profile\_length description\_length requirements\_length benefits\_length industry\_SWOE function\_SWOE from\_US money\_in\_title mentions\_salary telecommuting has\_company\_logo has\_questions has\_email has\_phone has\_url has\_fraud\_bigram has\_legit\_bigram;

/\* Determine population proportion of fraudulent job ads \*/

%global rho1;

**proc** **sql**;

select mean(fraudulent) into: rho1

from capstone.EMSCAD\_final;

**quit**;

/\* All-variables model \*/

**proc** **logistic** data=capstone.training;

class employment\_type(ref='Unspecified') required\_experience(ref='Unspecified')

required\_education(ref='Unspecified') / param=ref;

model fraudulent(event='1')=&all\_vars;

score data=capstone.validation priorevent=&rho1 fitstat

out=capstone.validation(rename=(p\_1=p\_allvars)

drop=F\_fraudulent I\_fraudulent p\_0);

**run**;

In the CLASS statement of the LOGISTIC procedure, the categorical variables *employment\_type*, *required\_experience*, and *required\_education* were encoded using reference cell encoding with the “Unspecified” category serving as the reference level for all three variables. The SCORE statement was used to score the validation dataset, storing the predicted probabilities of the job advertisements being fraudulent as a new column called “p\_allvars”. The PRIOREVENT option in the SCORE statement adjusts the predicted probabilities for oversampling using , the proportion of fraudulent job advertisements in the initial EMSCAD\_final dataset. The FITSTAT option displays various model fit statistics for assessing the predictive power of the all-variables model. The primary outputs of the LOGISTIC procedure for the all-variables model are shown in Tables 9-13.

**Table 9:** Hypothesis testing (all-variables model).[[4]](#footnote-4)

| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| --- | --- | --- | --- |
| Likelihood Ratio | 1382.5445 | 38 | <.0001 |
| Score | 1070.9115 | 38 | <.0001 |
| Wald | 485.4209 | 38 | <.0001 |

**Table 10:** Type 3 analysis of effects (all-variables model).

| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| --- | --- | --- | --- |
| employment\_type | 5 | 3.3435 | 0.6472 |
| required\_experience | 7 | 4.8834 | 0.6742 |
| required\_education | 9 | 21.5566 | 0.0104 |
| company\_profile\_length | 1 | 28.4588 | <.0001 |
| description\_length | 1 | 3.9137 | 0.0479 |
| requirements\_length | 1 | 4.2647 | 0.0389 |
| benefits\_length | 1 | 10.4247 | 0.0012 |
| industry\_SWOE | 1 | 113.2490 | <.0001 |
| function\_SWOE | 1 | 34.1659 | <.0001 |
| from\_US | 1 | 37.1678 | <.0001 |
| money\_in\_title | 1 | 17.3059 | <.0001 |
| mentions\_salary | 1 | 31.0755 | <.0001 |
| telecommuting | 1 | 0.0480 | 0.8265 |
| has\_company\_logo | 1 | 51.4228 | <.0001 |
| has\_questions | 1 | 1.5341 | 0.2155 |
| has\_email | 1 | 5.8718 | 0.0154 |
| has\_phone | 1 | 6.1146 | 0.0134 |
| has\_url | 1 | 0.7015 | 0.4023 |
| has\_fraud\_bigram | 1 | 73.7307 | <.0001 |
| has\_legit\_bigram | 1 | 8.7629 | 0.0031 |

**Table 11:** Analysis of maximum likelihood estimates (all-variables model).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Category** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept |  | 1 | 3.2745 | 0.4156 | 62.0851 | <.0001 |
| employment\_type | Contract | 1 | -0.0128 | 0.3314 | 0.0015 | 0.9693 |
| employment\_type | Full-time | 1 | -0.2742 | 0.2140 | 1.6417 | 0.2001 |
| employment\_type | Other | 1 | 0.4946 | 0.6276 | 0.6209 | 0.4307 |
| employment\_type | Part-time | 1 | -0.1495 | 0.3181 | 0.2210 | 0.6382 |
| employment\_type | Temporary | 1 | -13.9709 | 705.3 | 0.0004 | 0.9842 |
| required\_experience | Associate | 1 | -0.5587 | 0.3461 | 2.6060 | 0.1065 |
| required\_experience | Director | 1 | 0.0618 | 0.5681 | 0.0118 | 0.9134 |
| required\_experience | Entry level | 1 | -0.3128 | 0.2852 | 1.2029 | 0.2727 |
| required\_experience | Executive | 1 | -0.2805 | 0.6909 | 0.1648 | 0.6848 |
| required\_experience | Internship | 1 | -0.3356 | 0.7791 | 0.1856 | 0.6666 |
| required\_experience | Mid-Senior level | 1 | -0.4958 | 0.2786 | 3.1666 | 0.0752 |
| required\_experience | Not Applicable | 1 | -0.1944 | 0.3303 | 0.3466 | 0.5561 |
| required\_education | Associate Degree | 1 | 1.1902 | 0.7970 | 2.2302 | 0.1353 |
| required\_education | Bachelor’s Degree | 1 | 0.7021 | 0.2612 | 7.2274 | 0.0072 |
| required\_education | Certification | 1 | 0.4298 | 0.6828 | 0.3962 | 0.5290 |
| required\_education | Doctorate | 1 | -12.6735 | 2770.9 | 0.0000 | 0.9964 |
| required\_education | High School or equivalent | 1 | 0.5531 | 0.2649 | 4.3604 | 0.0368 |
| required\_education | Master’s Degree | 1 | 1.9047 | 0.4856 | 15.3881 | <.0001 |
| required\_education | Professional | 1 | -0.5885 | 1.4488 | 0.1650 | 0.6846 |
| required\_education | Some college coursework | 1 | 1.5433 | 1.0009 | 2.3772 | 0.1231 |
| required\_education | Vocational | 1 | -15.1643 | 1062.9 | 0.0002 | 0.9886 |
| company\_profile\_length |  | 1 | -0.00888 | 0.00166 | 28.4588 | <.0001 |
| description\_length |  | 1 | -0.00106 | 0.000535 | 3.9137 | 0.0479 |
| requirements\_length |  | 1 | -0.00190 | 0.000918 | 4.2647 | 0.0389 |
| benefits\_length |  | 1 | 0.00483 | 0.00150 | 10.4247 | 0.0012 |
| industry\_SWOE |  | 1 | 0.8397 | 0.0789 | 113.2490 | <.0001 |
| function\_SWOE |  | 1 | 0.5868 | 0.1004 | 34.1659 | <.0001 |
| from\_US |  | 1 | 1.1358 | 0.1863 | 37.1678 | <.0001 |
| money\_in\_title |  | 1 | 2.0637 | 0.4961 | 17.3059 | <.0001 |
| mentions\_salary |  | 1 | 1.1383 | 0.2042 | 31.0755 | <.0001 |
| telecommuting |  | 1 | 0.0662 | 0.3023 | 0.0480 | 0.8265 |
| has\_company\_logo |  | 1 | -1.4743 | 0.2056 | 51.4228 | <.0001 |
| has\_questions |  | 1 | -0.2008 | 0.1621 | 1.5341 | 0.2155 |
| has\_email |  | 1 | 0.7775 | 0.3209 | 5.8718 | 0.0154 |
| has\_phone |  | 1 | 0.9291 | 0.3757 | 6.1146 | 0.0134 |
| has\_url |  | 1 | 0.1387 | 0.1656 | 0.7015 | 0.4023 |
| has\_fraud\_bigram |  | 1 | 1.3820 | 0.1609 | 73.7307 | <.0001 |
| has\_legit\_bigram |  | 1 | -0.4379 | 0.1479 | 8.7629 | 0.0031 |

**Table 12:** Odds ratio estimates (all-variables model).

|  |  |  |  |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| employment\_type Contract vs. Unspecified | 0.987 | 0.516 | 1.890 |
| employment\_type Full-time vs. Unspecified | 0.760 | 0.500 | 1.156 |
| employment\_type Other vs. Unspecified | 1.640 | 0.479 | 5.611 |
| employment\_type Part-time vs. Unspecified | 0.861 | 0.462 | 1.606 |
| employment\_type Temporary vs. Unspecified | <0.001 | <0.001 | >999.999 |
| required\_experience Associate vs. Unspecified | 0.572 | 0.290 | 1.127 |
| required\_experience Director vs. Unspecified | 1.064 | 0.349 | 3.239 |
| required\_experience Entry level vs. Unspecified | 0.731 | 0.418 | 1.279 |
| required\_experience Executive vs. Unspecified | 0.755 | 0.195 | 2.926 |
| required\_experience Internship vs. Unspecified | 0.715 | 0.155 | 3.292 |
| required\_experience Mid-Senior level vs. Unspecified | 0.609 | 0.353 | 1.052 |
| required\_experience Not Applicable vs. Unspecified | 0.823 | 0.431 | 1.573 |
| required\_education Associate Degree vs. Unspecified | 3.288 | 0.689 | 15.678 |
| required\_education Bachelor’s Degree vs. Unspecified | 2.018 | 1.210 | 3.367 |
| required\_education Certification vs. Unspecified | 1.537 | 0.403 | 5.860 |
| required\_education Doctorate vs. Unspecified | <0.001 | <0.001 | >999.999 |
| required\_education High School or equivalent vs. Unspecified | 1.739 | 1.035 | 2.922 |
| required\_education Master’s Degree vs. Unspecified | 6.718 | 2.594 | 17.399 |
| required\_education Professional vs. Unspecified | 0.555 | 0.032 | 9.497 |
| required\_education Some college coursework vs. Unspecified | 4.680 | 0.658 | 33.283 |
| required\_education Vocational vs. Unspecified | <0.001 | <0.001 | >999.999 |
| company\_profile\_length | 0.991 | 0.988 | 0.994 |
| description\_length | 0.999 | 0.998 | 1.000 |
| requirements\_length | 0.998 | 0.996 | 1.000 |
| benefits\_length | 1.005 | 1.002 | 1.008 |
| industry\_SWOE | 2.316 | 1.984 | 2.703 |
| function\_SWOE | 1.798 | 1.477 | 2.189 |
| from\_US | 3.114 | 2.161 | 4.486 |
| money\_in\_title | 7.875 | 2.978 | 20.821 |
| mentions\_salary | 3.122 | 2.092 | 4.658 |
| telecommuting | 1.068 | 0.591 | 1.932 |
| has\_company\_logo | 0.229 | 0.153 | 0.343 |
| has\_questions | 0.818 | 0.595 | 1.124 |
| has\_email | 2.176 | 1.160 | 4.081 |
| has\_phone | 2.532 | 1.212 | 5.288 |
| has\_url | 1.149 | 0.830 | 1.589 |
| has\_fraud\_bigram | 3.983 | 2.905 | 5.460 |
| has\_legit\_bigram | 0.645 | 0.483 | 0.862 |

**Table 13:** Model fit statistics for validation dataset (all-variables model).

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Log Likelihood | -463.4 |
| Error Rate | 0.1753 |
| AIC | 1004.811 |
| AICC | 1007.937 |
| BIC | 1197.668 |
| SC | 1197.668 |
| R-Square | 0.499674 |
| Max-Rescaled R-Square | 0.628419 |
| AUC | 0.925456 |
| Brier Score | 0.127046 |

**Stepwise Selection Model**

The second logistic regression model that was explored was a stepwise selection model where SAS sequentially adds or removes variables from the model until only variables that are statistically significant at a predetermined significance level remain. Here is the code that was used to create the stepwise selection model:

/\* Stepwise selection model \*/

**proc** **logistic** data=capstone.training;

class employment\_type(ref='Unspecified') required\_experience(ref='Unspecified')

required\_education(ref='Unspecified') / param=ref;

model fraudulent(event='1')=&all\_vars / selection=stepwise;

score data=capstone.validation priorevent=&rho1 fitstat

out=capstone.validation(rename=(p\_1=p\_stepwise)

drop=F\_fraudulent I\_fraudulent p\_0);

**run**;

The LOGISTIC procedure shown here is similar to the one that was used to create the all-variables model except that the SELECTION=stepwise option was added to the MODEL statement to perform stepwise selection at the default 5% significance level. Again, a SCORE statement was included to score the validation dataset, storing the predicted probabilities of a job advertisement being fraudulent (corrected for oversampling using the PRIOREVENT option) as a new column called “p\_stepwise.” The primary outputs of the LOGISTIC procedure for the stepwise selection model are shown in Tables 14-18.

**Table 14:** Summary of stepwise selection.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Entered** | **Removed** |
| **1** | has\_company\_logo |  | 1 | 1 | 536.4143 |  | <.0001 |
| **2** | has\_fraud\_bigram |  | 1 | 2 | 311.4560 |  | <.0001 |
| **3** | industry\_SWOE |  | 1 | 3 | 176.7411 |  | <.0001 |
| **4** | function\_SWOE |  | 1 | 4 | 42.5643 |  | <.0001 |
| **5** | has\_phone |  | 1 | 5 | 32.7014 |  | <.0001 |
| **6** | company\_profile\_length |  | 1 | 6 | 32.6060 |  | <.0001 |
| **7** | mentions\_salary |  | 1 | 7 | 32.7504 |  | <.0001 |
| **8** | from\_US |  | 1 | 8 | 33.9760 |  | <.0001 |
| **9** | money\_in\_title |  | 1 | 9 | 26.8348 |  | <.0001 |
| **10** | benefits\_length |  | 1 | 10 | 15.7599 |  | <.0001 |
| **11** | has\_legit\_bigram |  | 1 | 11 | 14.5310 |  | 0.0001 |
| **12** | required\_education |  | 9 | 12 | 21.4043 |  | 0.0110 |
| **13** | requirements\_length |  | 1 | 13 | 6.1172 |  | 0.0134 |
| **14** | has\_email |  | 1 | 14 | 6.1185 |  | 0.0134 |
| **15** | description\_length |  | 1 | 15 | 3.9300 |  | 0.0474 |

**Table 15:** Type 3 analysis of effects (stepwise selection model).

|  |  |  |  |
| --- | --- | --- | --- |
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| required\_education | 9 | 18.6373 | 0.0285 |
| company\_profile\_length | 1 | 31.6527 | <.0001 |
| description\_length | 1 | 3.8814 | 0.0488 |
| requirements\_length | 1 | 5.4356 | 0.0197 |
| benefits\_length | 1 | 9.4155 | 0.0022 |
| industry\_SWOE | 1 | 117.6376 | <.0001 |
| function\_SWOE | 1 | 37.0939 | <.0001 |
| from\_US | 1 | 38.5176 | <.0001 |
| money\_in\_title | 1 | 18.8402 | <.0001 |
| mentions\_salary | 1 | 28.0055 | <.0001 |
| has\_company\_logo | 1 | 59.3067 | <.0001 |
| has\_email | 1 | 6.0338 | 0.0140 |
| has\_phone | 1 | 5.4669 | 0.0194 |
| has\_fraud\_bigram | 1 | 83.9056 | <.0001 |
| has\_legit\_bigram | 1 | 9.9812 | 0.0016 |

**Table 16:** 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 |

**Table 17:** Odds ratio estimates (stepwise selection model).

| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| --- | --- | --- | --- |
| required\_education Associate Degree vs. Unspecified | 2.041 | 0.455 | 9.159 |
| required\_education Bachelor's Degree vs. Unspecified | 1.433 | 0.934 | 2.199 |
| required\_education Certification vs. Unspecified | 1.238 | 0.338 | 4.540 |
| required\_education Doctorate vs. Unspecified | <0.001 | <0.001 | >999.999 |
| required\_education High School or equivalent vs. Unspecified | 1.304 | 0.857 | 1.984 |
| required\_education Master's Degree vs. Unspecified | 5.152 | 2.279 | 11.647 |
| required\_education Professional vs. Unspecified | 0.364 | 0.024 | 5.475 |
| required\_education Some college coursework vs. Unspecified | 2.801 | 0.415 | 18.887 |
| required\_education Vocational vs. Unspecified | <0.001 | <0.001 | >999.999 |
| company\_profile\_length | 0.991 | 0.988 | 0.994 |
| description\_length | 0.999 | 0.998 | 1.000 |
| requirements\_length | 0.998 | 0.996 | 1.000 |
| benefits\_length | 1.004 | 1.002 | 1.007 |
| industry\_SWOE | 2.308 | 1.984 | 2.685 |
| function\_SWOE | 1.796 | 1.488 | 2.168 |
| from\_US | 3.050 | 2.145 | 4.337 |
| money\_in\_title | 7.460 | 3.011 | 18.485 |
| mentions\_salary | 2.714 | 1.875 | 3.928 |
| has\_company\_logo | 0.214 | 0.144 | 0.316 |
| has\_email | 2.181 | 1.171 | 4.062 |
| has\_phone | 2.367 | 1.150 | 4.875 |
| has\_fraud\_bigram | 4.066 | 3.012 | 5.490 |
| has\_legit\_bigram | 0.634 | 0.477 | 0.841 |

**Table 18:** Model fit statistics for validation dataset (stepwise selection model).

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Log Likelihood | -440.6 |
| Error Rate | 0.1744 |
| AIC | 929.269 |
| AICC | 930.4536 |
| BIC | 1047.95 |
| SC | 1047.95 |
| R-Square | 0.521151 |
| Max-Rescaled R-Square | 0.655429 |
| AUC | 0.932009 |
| Brier Score | 0.12791 |

**Comparing the Models**

The last stage of the analysis was to compare the all-variables model to the stepwise selection model and pick the best model. First, looking at Tables 13 and 18, the stepwise selection model performs better than the all-variables model according to most of the model fit statistics. For example, when scoring the validation dataset, the all-variables model had an error rate of 17.53% while the stepwise selection model had an error rate of 17.44%, a small improvement. Also, the stepwise selection model has the advantage that it is more parsimonious (15 predictor variables) than the all-variables model (20 predictor variables). In general, given a group of similar models, the least complex model is preferred (Patetta, Lesson 4.1). Finally, the ROC curves of the two models were compared using the following code:

/\* Compare ROC curves \*/

ods select ROCOverlay;

**proc** **logistic** data=capstone.validation;

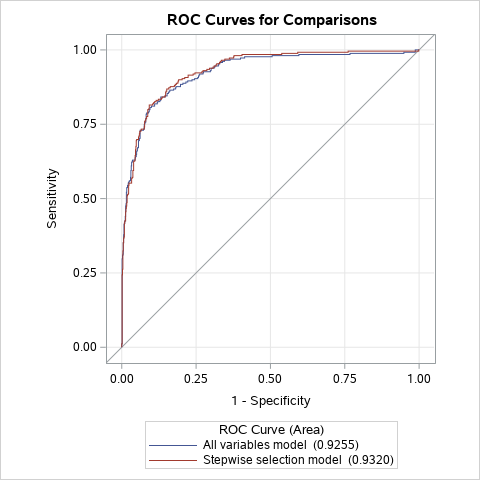
model fraudulent(event='1')=p\_allvars p\_stepwise / nofit;

roc 'All variables model' p\_allvars;

roc 'Stepwise selection model' p\_stepwise;

**run**;

**Figure 23:** ROC curve comparison for all-variables model and stepwise selection model.



In Figure 23, the ROC curves of the all-variables model and the stepwise selection model look very similar. However, technically speaking, the stepwise selection model does have a slightly higher AUC (0.9320) than the all-variables model (0.9255), so once again, the stepwise selection model is superior in this regard. In the end, after comparing the models with several different metrics, the stepwise selection model was chosen as the best model.

Now that a final model has been decided on, let us examine some of its features in more detail. From Table 15, the most statistically significant predictors of fraudulent status in the stepwise selection model are *company\_profile\_length*, *industry\_SWOE*, *function\_SWOE*, *from\_US*, *money\_in\_title*, *mentions\_salary*, *has\_company\_logo*, and *has\_fraud\_bigram* which all have Wald Test p-values below 0.001. Out of the three categorical variables *employment\_type*, *required\_experience*, and *required\_education* in the EMSCAD\_final dataset, only *required\_education* made it into the stepwise selection model. In contrast, all four of the text feature length variables *company\_profile\_length*, *description\_length*, *requirements\_length*, and *benefits\_length* were found to be significant at the 5% level.

According to Table 17, the binary variables which have the highest odds ratios are *from\_US*, *has\_fraud\_bigram*, and *money\_in\_title*. In particular, *from\_US* has an odds ratio of 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. Similarly, 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 (see pg. 25). Most impressively, 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. Also noteworthy is the odds ratio for *has\_company\_logo* which 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. Likewise, 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.

CONCLUSION

**Summary of Findings**

In this study, the research question was, “Which features of a job advertisement can help identify whether the advertisement is fraudulent?” To answer the research question, a logistic regression model was developed using the Employment Scam Aegean Dataset, a publicly available dataset of 17,880 online job advertisements that were classified as legitimate or fraudulent by researchers at the University of the Aegean. The null hypothesis stated that there is no statistically significant association between the job advertisement features in the study and the probability of an advertisement being fraudulent while the alternative hypothesis stated that there is a statistically significant association between at least one of the job advertisement features in the study and the probability of an advertisement being fraudulent. The stepwise selection model that was chosen as the best model contains 15 job advertisement features as predictors of fraudulent status which are all significant at the 5% significance level. Hence, we reject in favor of . 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*) are highly significant with Wald Test p-values below 0.001. Overall, the stepwise selection model scored the validation dataset with 82.56% accuracy.

What are the main takeaways of the study? To avoid becoming the victim of a job advertisement scam, the recommended course of action is to follow the six guidelines below 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% (Figure 13). 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 (Figures 8-11). 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!” In the Text Mining section, 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 (see pg. 25 for full list), 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 (Figures 16 & 17). 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 (Figures 5-7). The categorical variable *required\_education* was identified by SAS 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 (Figures 18 & 19). 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.

Following these six guidelines can help the general public avoid job advertisement scams and the problems associated with them such as financial loss, identity theft, and damaged reputations.

**Limitations**

There were several limitations to the study which should be addressed. First, 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. Another limitation of the study is that 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 results of the Text Mining section. Finally, the job advertisements in the EMSCAD were released 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.

**Further Research**

There are many ways in which the study could be expanded upon. First, a different predictive modelling technique could be used besides logistic regression. Past studies such as (Vidros et al., 2017) and (Kumar, 2020) have experimented with other models such as decision trees, random forests, and support vector machines with varying degrees of success. Another avenue of further research is to focus on job advertisements from one particular industry. It would be interesting to see what kinds of differences exist (if any) between fraudulent job advertisements from, say, the Information Technology sector compared to Sales & Marketing positions. Finally, as mentioned in the Limitations section, it might be beneficial to examine job advertisements from more recent years than the 2012-2014 range or job advertisements written in other languages besides English to see if the results of the study generalize to a broader set of job advertisements.

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>

Bhalla, Deepanshu. (2015, April). *Oversampling for Rare Event*. Listen Data. <https://www.listendata.com/2015/04/oversampling-for-rare-event.html>

Bird, S., Klein, E. & Loper, E. (2009). Natural Language Processing with Python. O’Reilly Media, Inc.

Goled, Shraddha. (2020, December 17). *Will SAS Language Continue To Hold Ground In Data Science?* Analytics India Magazine. <https://analyticsindiamag.com/will-sas-continue-to-hold-ground-in-data-science/>

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/>

Matange, Sanjay. (2015, December 23). *Box Plot with Stat Table and Markers*. SAS Blogs. <https://blogs.sas.com/content/graphicallyspeaking/2015/12/23/box-plot-with-stat-table-and-markers/>

Maurer, Roy. (2015, December 21). *Online Job Searching Has Doubled Since 2005*. Society for Human Resource Management. <https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/online-job-searching-doubled.aspx>

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>

Sivarajah, Sivakar. (2020, June 12). *“Sklearn’s TF-IDF” vs. “Standard TF-IDF”*. Towards Data Science. <https://towardsdatascience.com/how-sklearns-tf-idf-is-different-from-the-standard-tf-idf-275fa582e73d>

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>

Wicklin, Rick. (2011, September 19). *Count the number of missing values for each variable.* SAS Blogs. <https://blogs.sas.com/content/iml/2011/09/19/count-the-number-of-missing-values-for-each-variable.html>

APPENDIX

**Table A:** Cross tabulation of *industry* with *fraudulent*.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | Frequency  Percent  Row %  Column % | | | **industry** | **fraudulent** | | | | --- | --- | --- | --- | | **0** | **1** | **Total** | | **Unspecified** | 4628  25.88  94.39  27.20 | 275  1.54  5.61  31.76 | 4903  27.42 | | **Information Technology and Services** | 1702  9.52  98.15  10.00 | 32  0.18  1.85  3.70 | 1734  9.70 | | **Computer Software** | 1371  7.67  99.64  8.06 | 5  0.03  0.36  0.58 | 1376  7.70 | | **Internet** | 1062  5.94  100.00  6.24 | 0  0.00  0.00  0.00 | 1062  5.94 | | **Marketing and Advertising** | 783  4.38  94.57  4.60 | 45  0.25  5.43  5.20 | 828  4.63 | | **Education Management** | 822  4.60  100.00  4.83 | 0  0.00  0.00  0.00 | 822  4.60 | | **Financial Services** | 744  4.16  95.51  4.37 | 35  0.20  4.49  4.04 | 779  4.36 | | **Hospital & Health Care** | 446  2.49  89.74  2.62 | 51  0.29  10.26  5.89 | 497  2.78 | | **Consumer Services** | 334  1.87  93.30  1.96 | 24  0.13  6.70  2.77 | 358  2.00 | | **Telecommunications** | 316  1.77  92.40  1.86 | 26  0.15  7.60  3.00 | 342  1.91 | | **Oil & Energy** | 178  1.00  62.02  1.05 | 109  0.61  37.98  12.59 | 287  1.61 | | **Retail** | 218  1.22  97.76  1.28 | 5  0.03  2.24  0.58 | 223  1.25 | | **Real Estate** | 151  0.84  86.29  0.89 | 24  0.13  13.71  2.77 | 175  0.98 | | **Accounting** | 102  0.57  64.15  0.60 | 57  0.32  35.85  6.58 | 159  0.89 | | **Construction** | 155  0.87  98.10  0.91 | 3  0.02  1.90  0.35 | 158  0.88 | | **E-Learning** | 137  0.77  98.56  0.81 | 2  0.01  1.44  0.23 | 139  0.78 | | **Management Consulting** | 124  0.69  95.38  0.73 | 6  0.03  4.62  0.69 | 130  0.73 | | **Design** | 125  0.70  96.90  0.73 | 4  0.02  3.10  0.46 | 129  0.72 | | **Health, Wellness and Fitness** | 112  0.63  88.19  0.66 | 15  0.08  11.81  1.73 | 127  0.71 | | **Staffing and Recruiting** | 119  0.67  93.70  0.70 | 8  0.04  6.30  0.92 | 127  0.71 | | **Insurance** | 117  0.65  95.12  0.69 | 6  0.03  4.88  0.69 | 123  0.69 | | **Automotive** | 115  0.64  95.83  0.68 | 5  0.03  4.17  0.58 | 120  0.67 | | **Logistics and Supply Chain** | 110  0.62  98.21  0.65 | 2  0.01  1.79  0.23 | 112  0.63 | | **Human Resources** | 102  0.57  94.44  0.60 | 6  0.03  5.56  0.69 | 108  0.60 | | **Online Media** | 100  0.56  99.01  0.59 | 1  0.01  0.99  0.12 | 101  0.56 | | **Apparel & Fashion** | 95  0.53  97.94  0.56 | 2  0.01  2.06  0.23 | 97  0.54 | | **Legal Services** | 97  0.54  100.00  0.57 | 0  0.00  0.00  0.00 | 97  0.54 | | **Facilities Services** | 92  0.51  97.87  0.54 | 2  0.01  2.13  0.23 | 94  0.53 | | **Hospitality** | 74  0.41  84.09  0.43 | 14  0.08  15.91  1.62 | 88  0.49 | | **Computer Games** | 86  0.48  100.00  0.51 | 0  0.00  0.00  0.00 | 86  0.48 | | **Banking** | 81  0.45  96.43  0.48 | 3  0.02  3.57  0.35 | 84  0.47 | | **Building Materials** | 77  0.43  98.72  0.45 | 1  0.01  1.28  0.12 | 78  0.44 | | **Leisure, Travel & Tourism** | 55  0.31  72.37  0.32 | 21  0.12  27.63  2.42 | 76  0.43 | | **Nonprofit Organization Management** | 76  0.43  100.00  0.45 | 0  0.00  0.00  0.00 | 76  0.43 | | **Entertainment** | 69  0.39  93.24  0.41 | 5  0.03  6.76  0.58 | 74  0.41 | | **Electrical/Electronic Manufacturing** | 69  0.39  94.52  0.41 | 4  0.02  5.48  0.46 | 73  0.41 | | **Food & Beverages** | 72  0.40  100.00  0.42 | 0  0.00  0.00  0.00 | 72  0.40 | | **Cosmetics** | 64  0.36  98.46  0.38 | 1  0.01  1.54  0.12 | 65  0.36 | | **Airlines/Aviation** | 62  0.35  98.41  0.36 | 1  0.01  1.59  0.12 | 63  0.35 | | **Consumer Goods** | 62  0.35  98.41  0.36 | 1  0.01  1.59  0.12 | 63  0.35 | | **Consumer Electronics** | 62  0.35  100.00  0.36 | 0  0.00  0.00  0.00 | 62  0.35 | | **Medical Practice** | 59  0.33  98.33  0.35 | 1  0.01  1.67  0.12 | 60  0.34 | | **Public Relations and Communications** | 58  0.32  100.00  0.34 | 0  0.00  0.00  0.00 | 58  0.32 | | **Civic & Social Organization** | 54  0.30  98.18  0.32 | 1  0.01  1.82  0.12 | 55  0.31 | | **Market Research** | 52  0.29  96.30  0.31 | 2  0.01  3.70  0.23 | 54  0.30 | | **Transportation/Trucking/Railroad** | 50  0.28  94.34  0.29 | 3  0.02  5.66  0.35 | 53  0.30 | | **Restaurants** | 52  0.29  100.00  0.31 | 0  0.00  0.00  0.00 | 52  0.29 | | **Warehousing** | 50  0.28  98.04  0.29 | 1  0.01  1.96  0.12 | 51  0.29 | | **Broadcast Media** | 49  0.27  98.00  0.29 | 1  0.01  2.00  0.12 | 50  0.28 | | **Events Services** | 50  0.28  100.00  0.29 | 0  0.00  0.00  0.00 | 50  0.28 | | **Computer & Network Security** | 44  0.25  89.80  0.26 | 5  0.03  10.20  0.58 | 49  0.27 | | **Environmental Services** | 46  0.26  93.88  0.27 | 3  0.02  6.12  0.35 | 49  0.27 | | **Media Production** | 45  0.25  93.75  0.26 | 3  0.02  6.25  0.35 | 48  0.27 | | **Computer Networking** | 32  0.18  72.73  0.19 | 12  0.07  27.27  1.39 | 44  0.25 | | **Food Production** | 43  0.24  97.73  0.25 | 1  0.01  2.27  0.12 | 44  0.25 | | **Gambling & Casinos** | 42  0.23  100.00  0.25 | 0  0.00  0.00  0.00 | 42  0.23 | | **Pharmaceuticals** | 42  0.23  100.00  0.25 | 0  0.00  0.00  0.00 | 42  0.23 | | **Publishing** | 39  0.22  100.00  0.23 | 0  0.00  0.00  0.00 | 39  0.22 | | **Biotechnology** | 34  0.19  89.47  0.20 | 4  0.02  10.53  0.46 | 38  0.21 | | **Mechanical or Industrial Engineering** | 33  0.18  89.19  0.19 | 4  0.02  10.81  0.46 | 37  0.21 | | **Computer Hardware** | 32  0.18  91.43  0.19 | 3  0.02  8.57  0.35 | 35  0.20 | | **Utilities** | 32  0.18  96.97  0.19 | 1  0.01  3.03  0.12 | 33  0.18 | | **Graphic Design** | 32  0.18  100.00  0.19 | 0  0.00  0.00  0.00 | 32  0.18 | | **Printing** | 30  0.17  100.00  0.18 | 0  0.00  0.00  0.00 | 30  0.17 | | **Security and Investigations** | 29  0.16  96.67  0.17 | 1  0.01  3.33  0.12 | 30  0.17 | | **Research** | 29  0.16  100.00  0.17 | 0  0.00  0.00  0.00 | 29  0.16 | | **Venture Capital & Private Equity** | 29  0.16  100.00  0.17 | 0  0.00  0.00  0.00 | 29  0.16 | | **Information Services** | 26  0.15  92.86  0.15 | 2  0.01  7.14  0.23 | 28  0.16 | | **Aviation & Aerospace** | 24  0.13  100.00  0.14 | 0  0.00  0.00  0.00 | 24  0.13 | | **Farming** | 24  0.13  100.00  0.14 | 0  0.00  0.00  0.00 | 24  0.13 | | **Mental Health Care** | 23  0.13  100.00  0.14 | 0  0.00  0.00  0.00 | 23  0.13 | | **Sports** | 23  0.13  100.00  0.14 | 0  0.00  0.00  0.00 | 23  0.13 | | **Chemicals** | 22  0.12  100.00  0.13 | 0  0.00  0.00  0.00 | 22  0.12 | | **Government Administration** | 22  0.12  100.00  0.13 | 0  0.00  0.00  0.00 | 22  0.12 | | **Law Practice** | 19  0.11  100.00  0.11 | 0  0.00  0.00  0.00 | 19  0.11 | | **Medical Devices** | 18  0.10  94.74  0.11 | 1  0.01  5.26  0.12 | 19  0.11 | | **Outsourcing/Offshoring** | 18  0.10  94.74  0.11 | 1  0.01  5.26  0.12 | 19  0.11 | | **Writing and Editing** | 19  0.11  100.00  0.11 | 0  0.00  0.00  0.00 | 19  0.11 | | **Business Supplies and Equipment** | 15  0.08  83.33  0.09 | 3  0.02  16.67  0.35 | 18  0.10 | | **Fund-Raising** | 16  0.09  100.00  0.09 | 0  0.00  0.00  0.00 | 16  0.09 | | **Professional Training & Coaching** | 14  0.08  100.00  0.08 | 0  0.00  0.00  0.00 | 14  0.08 | | **Government Relations** | 11  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 11  0.06 | | **Higher Education** | 11  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 11  0.06 | | **Machinery** | 11  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 11  0.06 | | **Semiconductors** | 11  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 11  0.06 | | **Wholesale** | 10  0.06  90.91  0.06 | 1  0.01  9.09  0.12 | 11  0.06 | | **Architecture & Planning** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **Law Enforcement** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **Music** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **Translation and Localization** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **Civil Engineering** | 8  0.04  88.89  0.05 | 1  0.01  11.11  0.12 | 9  0.05 | | **Defense & Space** | 7  0.04  77.78  0.04 | 2  0.01  22.22  0.23 | 9  0.05 | | **Individual & Family Services** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **Program Development** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **Renewables & Environment** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **Executive Office** | 6  0.03  75.00  0.04 | 2  0.01  25.00  0.23 | 8  0.04 | | **International Trade and Development** | 8  0.04  100.00  0.05 | 0  0.00  0.00  0.00 | 8  0.04 | | **Veterinary** | 8  0.04  100.00  0.05 | 0  0.00  0.00  0.00 | 8  0.04 | | **Industrial Automation** | 7  0.04  100.00  0.04 | 0  0.00  0.00  0.00 | 7  0.04 | | **Photography** | 7  0.04  100.00  0.04 | 0  0.00  0.00  0.00 | 7  0.04 | | **Public Safety** | 6  0.03  85.71  0.04 | 1  0.01  14.29  0.12 | 7  0.04 | | **Investment Management** | 5  0.03  83.33  0.03 | 1  0.01  16.67  0.12 | 6  0.03 | | **Motion Pictures and Film** | 6  0.03  100.00  0.04 | 0  0.00  0.00  0.00 | 6  0.03 | | **Primary/Secondary Education** | 6  0.03  100.00  0.04 | 0  0.00  0.00  0.00 | 6  0.03 | | **Religious Institutions** | 6  0.03  100.00  0.04 | 0  0.00  0.00  0.00 | 6  0.03 | | **Animation** | 3  0.02  60.00  0.02 | 2  0.01  40.00  0.23 | 5  0.03 | | **Capital Markets** | 5  0.03  100.00  0.03 | 0  0.00  0.00  0.00 | 5  0.03 | | **Import and Export** | 5  0.03  100.00  0.03 | 0  0.00  0.00  0.00 | 5  0.03 | | **Package/Freight Delivery** | 5  0.03  100.00  0.03 | 0  0.00  0.00  0.00 | 5  0.03 | | **Packaging and Containers** | 5  0.03  100.00  0.03 | 0  0.00  0.00  0.00 | 5  0.03 | | **Commercial Real Estate** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Fishery** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Investment Banking** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Luxury Goods & Jewelry** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Philanthropy** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Wireless** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **Furniture** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Maritime** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Mining & Metals** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Performing Arts** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Plastics** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Public Policy** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **Libraries** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **Military** | 1  0.01  50.00  0.01 | 1  0.01  50.00  0.12 | 2  0.01 | | **Nanotechnology** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **Textiles** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **Alternative Dispute Resolution** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Museums and Institutions** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Ranching** | 0  0.00  0.00  0.00 | 1  0.01  100.00  0.12 | 1  0.01 | | **Shipbuilding** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Sporting Goods** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Wine and Spirits** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Total** | 17014  95.16 | 866  4.84 | 17880  100.00 | |

**Table B:** Cross tabulation of *function* with *fraudulent*.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | Frequency  Percent  Row %  Column % | | | **function** | **fraudulent** | | | | --- | --- | --- | --- | | **0** | **1** | **Total** | | **Unspecified** | 6118  34.22  94.78  35.96 | 337  1.88  5.22  38.91 | 6455  36.10 | | **Information Technology** | 1717  9.60  98.17  10.09 | 32  0.18  1.83  3.70 | 1749  9.78 | | **Sales** | 1427  7.98  97.21  8.39 | 41  0.23  2.79  4.73 | 1468  8.21 | | **Engineering** | 1235  6.91  91.62  7.26 | 113  0.63  8.38  13.05 | 1348  7.54 | | **Customer Service** | 1162  6.50  94.55  6.83 | 67  0.37  5.45  7.74 | 1229  6.87 | | **Marketing** | 820  4.59  98.80  4.82 | 10  0.06  1.20  1.15 | 830  4.64 | | **Administrative** | 511  2.86  81.11  3.00 | 119  0.67  18.89  13.74 | 630  3.52 | | **Design** | 337  1.88  99.12  1.98 | 3  0.02  0.88  0.35 | 340  1.90 | | **Health Care Provider** | 337  1.88  99.70  1.98 | 1  0.01  0.30  0.12 | 338  1.89 | | **Education** | 324  1.81  99.69  1.90 | 1  0.01  0.31  0.12 | 325  1.82 | | **Other** | 293  1.64  90.15  1.72 | 32  0.18  9.85  3.70 | 325  1.82 | | **Management** | 311  1.74  98.11  1.83 | 6  0.03  1.89  0.69 | 317  1.77 | | **Business Development** | 215  1.20  94.30  1.26 | 13  0.07  5.70  1.50 | 228  1.28 | | **Accounting/Auditing** | 183  1.02  86.32  1.08 | 29  0.16  13.68  3.35 | 212  1.19 | | **Human Resources** | 196  1.10  95.61  1.15 | 9  0.05  4.39  1.04 | 205  1.15 | | **Project Management** | 173  0.97  94.54  1.02 | 10  0.06  5.46  1.15 | 183  1.02 | | **Finance** | 157  0.88  91.28  0.92 | 15  0.08  8.72  1.73 | 172  0.96 | | **Consulting** | 140  0.78  97.22  0.82 | 4  0.02  2.78  0.46 | 144  0.81 | | **Art/Creative** | 131  0.73  99.24  0.77 | 1  0.01  0.76  0.12 | 132  0.74 | | **Writing/Editing** | 132  0.74  100.00  0.78 | 0  0.00  0.00  0.00 | 132  0.74 | | **Production** | 116  0.65  100.00  0.68 | 0  0.00  0.00  0.00 | 116  0.65 | | **Product Management** | 114  0.64  100.00  0.67 | 0  0.00  0.00  0.00 | 114  0.64 | | **Quality Assurance** | 111  0.62  100.00  0.65 | 0  0.00  0.00  0.00 | 111  0.62 | | **Advertising** | 85  0.48  94.44  0.50 | 5  0.03  5.56  0.58 | 90  0.50 | | **Business Analyst** | 83  0.46  98.81  0.49 | 1  0.01  1.19  0.12 | 84  0.47 | | **Data Analyst** | 78  0.44  95.12  0.46 | 4  0.02  4.88  0.46 | 82  0.46 | | **Public Relations** | 75  0.42  98.68  0.44 | 1  0.01  1.32  0.12 | 76  0.43 | | **Manufacturing** | 72  0.40  97.30  0.42 | 2  0.01  2.70  0.23 | 74  0.41 | | **General Business** | 67  0.37  98.53  0.39 | 1  0.01  1.47  0.12 | 68  0.38 | | **Research** | 50  0.28  100.00  0.29 | 0  0.00  0.00  0.00 | 50  0.28 | | **Legal** | 47  0.26  100.00  0.28 | 0  0.00  0.00  0.00 | 47  0.26 | | **Strategy/Planning** | 45  0.25  97.83  0.26 | 1  0.01  2.17  0.12 | 46  0.26 | | **Training** | 38  0.21  100.00  0.22 | 0  0.00  0.00  0.00 | 38  0.21 | | **Supply Chain** | 36  0.20  100.00  0.21 | 0  0.00  0.00  0.00 | 36  0.20 | | **Financial Analyst** | 28  0.16  84.85  0.16 | 5  0.03  15.15  0.58 | 33  0.18 | | **Distribution** | 21  0.12  87.50  0.12 | 3  0.02  12.50  0.35 | 24  0.13 | | **Purchasing** | 15  0.08  100.00  0.09 | 0  0.00  0.00  0.00 | 15  0.08 | | **Science** | 14  0.08  100.00  0.08 | 0  0.00  0.00  0.00 | 14  0.08 | | **Total** | 17014  95.16 | 866  4.84 | 17880  100.00 | |

**Table C:** Cross tabulation of *country* with *fraudulent*.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | Frequency  Percent  Row %  Column % | | | **country** | **fraudulent** | | | | --- | --- | --- | --- | | **0** | **1** | **Total** | | **US** | 9926  55.51  93.15  58.34 | 730  4.08  6.85  84.30 | 10656  59.60 | | **GB** | 2361  13.20  99.04  13.88 | 23  0.13  0.96  2.66 | 2384  13.33 | | **GR** | 940  5.26  100.00  5.52 | 0  0.00  0.00  0.00 | 940  5.26 | | **CA** | 445  2.49  97.37  2.62 | 12  0.07  2.63  1.39 | 457  2.56 | | **DE** | 383  2.14  100.00  2.25 | 0  0.00  0.00  0.00 | 383  2.14 | | **Unspecified** | 327  1.83  94.51  1.92 | 19  0.11  5.49  2.19 | 346  1.94 | | **NZ** | 333  1.86  100.00  1.96 | 0  0.00  0.00  0.00 | 333  1.86 | | **IN** | 272  1.52  98.55  1.60 | 4  0.02  1.45  0.46 | 276  1.54 | | **AU** | 174  0.97  81.31  1.02 | 40  0.22  18.69  4.62 | 214  1.20 | | **PH** | 131  0.73  99.24  0.77 | 1  0.01  0.76  0.12 | 132  0.74 | | **NL** | 127  0.71  100.00  0.75 | 0  0.00  0.00  0.00 | 127  0.71 | | **BE** | 117  0.65  100.00  0.69 | 0  0.00  0.00  0.00 | 117  0.65 | | **IE** | 114  0.64  100.00  0.67 | 0  0.00  0.00  0.00 | 114  0.64 | | **SG** | 80  0.45  100.00  0.47 | 0  0.00  0.00  0.00 | 80  0.45 | | **HK** | 77  0.43  100.00  0.45 | 0  0.00  0.00  0.00 | 77  0.43 | | **PL** | 73  0.41  96.05  0.43 | 3  0.02  3.95  0.35 | 76  0.43 | | **EE** | 71  0.40  98.61  0.42 | 1  0.01  1.39  0.12 | 72  0.40 | | **IL** | 72  0.40  100.00  0.42 | 0  0.00  0.00  0.00 | 72  0.40 | | **FR** | 70  0.39  100.00  0.41 | 0  0.00  0.00  0.00 | 70  0.39 | | **ES** | 65  0.36  98.48  0.38 | 1  0.01  1.52  0.12 | 66  0.37 | | **AE** | 53  0.30  98.15  0.31 | 1  0.01  1.85  0.12 | 54  0.30 | | **EG** | 51  0.29  98.08  0.30 | 1  0.01  1.92  0.12 | 52  0.29 | | **SE** | 49  0.27  100.00  0.29 | 0  0.00  0.00  0.00 | 49  0.27 | | **RO** | 46  0.26  100.00  0.27 | 0  0.00  0.00  0.00 | 46  0.26 | | **DK** | 42  0.23  100.00  0.25 | 0  0.00  0.00  0.00 | 42  0.23 | | **ZA** | 39  0.22  97.50  0.23 | 1  0.01  2.50  0.12 | 40  0.22 | | **BR** | 35  0.20  97.22  0.21 | 1  0.01  2.78  0.12 | 36  0.20 | | **IT** | 31  0.17  100.00  0.18 | 0  0.00  0.00  0.00 | 31  0.17 | | **FI** | 29  0.16  100.00  0.17 | 0  0.00  0.00  0.00 | 29  0.16 | | **PK** | 26  0.15  96.30  0.15 | 1  0.01  3.70  0.12 | 27  0.15 | | **LT** | 23  0.13  100.00  0.14 | 0  0.00  0.00  0.00 | 23  0.13 | | **MY** | 9  0.05  42.86  0.05 | 12  0.07  57.14  1.39 | 21  0.12 | | **QA** | 15  0.08  71.43  0.09 | 6  0.03  28.57  0.69 | 21  0.12 | | **JP** | 20  0.11  100.00  0.12 | 0  0.00  0.00  0.00 | 20  0.11 | | **RU** | 20  0.11  100.00  0.12 | 0  0.00  0.00  0.00 | 20  0.11 | | **MX** | 18  0.10  100.00  0.11 | 0  0.00  0.00  0.00 | 18  0.10 | | **PT** | 18  0.10  100.00  0.11 | 0  0.00  0.00  0.00 | 18  0.10 | | **BG** | 17  0.10  100.00  0.10 | 0  0.00  0.00  0.00 | 17  0.10 | | **TR** | 17  0.10  100.00  0.10 | 0  0.00  0.00  0.00 | 17  0.10 | | **CH** | 15  0.08  100.00  0.09 | 0  0.00  0.00  0.00 | 15  0.08 | | **CN** | 15  0.08  100.00  0.09 | 0  0.00  0.00  0.00 | 15  0.08 | | **SA** | 14  0.08  93.33  0.08 | 1  0.01  6.67  0.12 | 15  0.08 | | **AT** | 14  0.08  100.00  0.08 | 0  0.00  0.00  0.00 | 14  0.08 | | **HU** | 14  0.08  100.00  0.08 | 0  0.00  0.00  0.00 | 14  0.08 | | **MU** | 14  0.08  100.00  0.08 | 0  0.00  0.00  0.00 | 14  0.08 | | **ID** | 12  0.07  92.31  0.07 | 1  0.01  7.69  0.12 | 13  0.07 | | **MT** | 13  0.07  100.00  0.08 | 0  0.00  0.00  0.00 | 13  0.07 | | **UA** | 13  0.07  100.00  0.08 | 0  0.00  0.00  0.00 | 13  0.07 | | **CY** | 11  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 11  0.06 | | **IQ** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **KR** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **NG** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **TH** | 10  0.06  100.00  0.06 | 0  0.00  0.00  0.00 | 10  0.06 | | **AR** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **BH** | 4  0.02  44.44  0.02 | 5  0.03  55.56  0.58 | 9  0.05 | | **BY** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **LU** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **PA** | 9  0.05  100.00  0.05 | 0  0.00  0.00  0.00 | 9  0.05 | | **NO** | 8  0.04  100.00  0.05 | 0  0.00  0.00  0.00 | 8  0.04 | | **KE** | 7  0.04  100.00  0.04 | 0  0.00  0.00  0.00 | 7  0.04 | | **RS** | 7  0.04  100.00  0.04 | 0  0.00  0.00  0.00 | 7  0.04 | | **CZ** | 6  0.03  100.00  0.04 | 0  0.00  0.00  0.00 | 6  0.03 | | **LV** | 6  0.03  100.00  0.04 | 0  0.00  0.00  0.00 | 6  0.03 | | **NI** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **TT** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **TW** | 2  0.01  50.00  0.01 | 2  0.01  50.00  0.23 | 4  0.02 | | **VN** | 4  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 4  0.02 | | **VI** | 3  0.02  100.00  0.02 | 0  0.00  0.00  0.00 | 3  0.02 | | **AM** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **BD** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **CL** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **IS** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **KW** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **LK** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **SK** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **TN** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **ZM** | 2  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 2  0.01 | | **AL** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **CM** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **CO** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **GH** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **HR** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **JM** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **KH** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **KZ** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **MA** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **PE** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **SD** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **SI** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **SV** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **UG** | 1  0.01  100.00  0.01 | 0  0.00  0.00  0.00 | 1  0.01 | | **Total** | 17014  95.16 | 866  4.84 | 17880  100.00 | |

|  |  |
| --- | --- |
|  |  |

1. <http://emscad.samos.aegean.gr/> [↑](#footnote-ref-1)
2. <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction> [↑](#footnote-ref-2)
3. In general, the larger the value of , the more aggressively the data is smoothed (Patetta, Lesson 3.2).

   The choice of is relatively small and has a minimal effect on the smoothed log-odds values. [↑](#footnote-ref-3)
4. The hypotheses being tested are the null hypothesis and alternative hypothesis from the Hypothesis section (pg. 2). Table 9 indicates that we reject in favor of at the 5% significance level for all three of the tests shown (Likelihood Ratio, Score, and Wald). [↑](#footnote-ref-4)