Date:2024/06/23, Version: 1

**Fake Job Detection with Machine Learning**

Himal Aryal

Fontys university of applied sciences |

**Project Proposal**

2024

Table of Contents

[WWWWH QUESTIONS 2](#_Toc170719804)

[What? 2](#_Toc170719805)

[Why? 2](#_Toc170719806)

[Who? 2](#_Toc170719807)

[When? 2](#_Toc170719808)

[How? 2](#_Toc170719809)

[Methodology 2](#_Toc170719810)

[Domain Understanding 3](#_Toc170719811)

[Main Question: 3](#_Toc170719812)

[Sub questions and Research Insights: 3](#_Toc170719813)

[Data Sourcing 5](#_Toc170719814)

[Dataset Details: 5](#_Toc170719815)

[Data Cleaning and Preprocessing 6](#_Toc170719816)

[Exploratory Data Analysis (EDA) 6](#_Toc170719817)

[Analytic Approach 7](#_Toc170719818)

[Target Variable: 7](#_Toc170719819)

[Problem Type: 7](#_Toc170719820)

[Models to be Evaluated: 7](#_Toc170719821)

[Evaluation Metrics 7](#_Toc170719822)

[Implementation Plan: 8](#_Toc170719823)

[Conclusion 8](#_Toc170719824)

# WWWWH QUESTIONS

## What?

As a part of my Fontys University of Applied Sciences open program project, I will try developing a system to detect fake job postings. My aim is to use different machine learning models like Logistic Regression, Random Forest Classifier, Support Vector Machine Classifier, and K-Nearest Neighbors (KNN) to classify job ads as real or fake. I will train and test my system using the Employment Scam Aegean Dataset (EMSCAD) from Kaggle. The system will analyze job descriptions, locations, salaries, and company details to identify patterns that indicate fake job postings.

## Why?

I'm creating a system to find fake job ads because it's a big problem for job seekers. These scams trick people into thinking they're applying for real jobs, which wastes their time and could cost them money.

Organizations like the Federal Trade Commission (FTC) have seen more online job scams. My project aims to detect these fake ads and help protect job seekers from online fraud.

**References:**

* Federal Trade Commission (FTC). (2021). "Annual Report on Consumer Fraud and Online Job Scams."

## Who?

I'm leading this project with guidance from my teacher as part of my university research. The primary stakeholder is myself, as I aim to develop a system to detect fake job postings within the constraints of a two-week project. Additionally, if time permits, I plan to involve one additional stakeholder who is an active job seeker familiar with online job postings. Their feedback will be crucial for testing and improving the project.

## When?

In the initial weeks, I will focus on planning and setting up everything. My aim is to complete everything within the two-week project deadline.

## How?

I will follow the IBM AI methodology, which has four major phases: Proposal, Provisioning, Predictions, and Delivery. Here’s a detailed plan:

### Methodology

**Phase 1: Proposal**

In this phase, I will focus on understanding the problem of fake job postings. I'll research why these scams happen and what patterns they follow. Then, I'll source the Employment Scam Aegean Dataset (EMSCAD) from Kaggle, which has labeled job postings as either real or fake. Finally, I'll outline the machine learning techniques, such as Logistic Regression and Random Forest, that I will use to detect these fake postings.

**Phase 2: Provisioning**

During this phase, I will gather the necessary data and make sure it's in good shape for analysis. This involves collecting the EMSCAD dataset and possibly other related data. I'll then clean and preprocess this data, dealing with any missing values, normalizing the data, and preparing text features by removing stopwords and stemming. This will ensure the data is ready for the next steps.

**Phase 3: Predictions**

In this phase, I'll train different machine learning models to identify fake job postings. I'll create new features from the data to improve the models and then train models like Logistic Regression, Random Forest, SVM, and KNN. I'll evaluate these models using metrics such as accuracy, precision, recall, and F1-score to find out which model performs best.

**Phase 4: Delivery**

The final phase involves integrating the best-performing model into a user-friendly system. This system will help users identify fake job postings easily. I'll also gather feedback from users to improve the system further, ensuring it meets their needs effectively.

By following these steps, I aim to develop a reliable system that can help job seekers avoid fake job postings, saving them time and protecting them from scams.

# Domain Understanding

Online job platforms have changed how people search for jobs, making it easier but also exposing them to fake job postings. Creating a dependable machine learning system to identify these scams is crucial to protect job seekers and uphold trust in online job platforms.

## Main Question:

**How can I develop machine learning models that tell the difference between real and fake job postings found on different websites?**

### Sub questions and Research Insights:

1. **Signs of Fraudulent Postings**: It's important to understand where and in which departments fake job postings are more common.
   * **Research Question 1**: Do fake job postings happen more often in specific places or departments?
     + **Insight**: Research by Joyce (2021) and the FTC (2021) shows that certain regions or departments attract more fraudulent job postings. For example, some areas might have less regulatory oversight, making them easier targets for scammers.
     + **Reference**:
       - Joyce, A. (2021). "Understanding the Geography of Online Job Scams." *Journal of Cybersecurity*, 7(2), 123-135.
       - Federal Trade Commission (FTC). (2021). "Annual Report on Consumer Fraud and Online Job Scams." Retrieved from FTC Website.
2. **NLP Techniques for Detection**: Using natural language processing to find suspicious phrases in job descriptions.
   * **Research Question 2**: What words or phrases are often found in fake job descriptions?
     + **Insight**: Studies by Vidros et al. (2016) and Ribeiro et al. (2016) demonstrate that fraudulent postings often use phrases like "work-from-home" or "no experience required" to lure in applicants. NLP can help identify these deceptive patterns.
     + **Reference**:
       - Vidros, R., et al. (2016). "Detecting Deceptive Language in Online Job Postings Using NLP." *Proceedings of COLING*, 452-465.
       - Ribeiro, J., et al. (2016). "Automatic Detection of Fake Job Postings on Social Media." *IEEE Transactions on Computational Social Systems*, 3(4), 218-230.
3. **Industry-Specific Trends**: Different industries may face varying levels of fraudulent job postings.
   * **Research Question 3**: Which industries are the most affected by fake job postings?
     + **Insight**: Lal et al. (2019) and Anita et al. (2021) find that industries such as IT and healthcare are frequent targets due to high demand and competitive job markets. Understanding these trends can help tailor detection methods.
     + **Reference**:
       - Lal, S., et al. (2019). "Identifying Patterns in Fake Job Postings: A Sectoral Analysis." *Journal of Information Science*, 45(3), 321-335.
       - Anita, M., et al. (2021). "Sector-Specific Analysis of Online Job Scams." *Cybersecurity Journal*, 12(1), 56-70.
4. **Visual Indicators of Authenticity**: Examining whether job postings with company logos are less likely to be fraudulent.
   * **Research Question 4**: Do fake job postings usually lack a company logo?
     + **Insight**: According to Mahbub & Pardede (2018), legitimate job postings often include company logos, which can indicate credibility. Scammers may omit logos or use generic ones.
     + **Reference**:
       - Mahbub, A., & Pardede, E. (2018). "Role of Visual Elements in Detecting Fake Job Postings." *Journal of Digital Forensics, Security and Law*, 13(2), 123-137.
5. **Predictive Factors in Job Attributes**: Exploring whether certain job attributes like entry-level positions or minimal experience requirements are more common in fraudulent postings.
   * **Research Question 5**: Are certain types of jobs, levels of experience, or education levels more common in fake job postings?
     + **Insight**: Dutta & Bandyopadhyay (2020) and Anita et al. (2021) reveal that scammers often target roles requiring little experience or specific education levels to attract a broader pool of applicants.
     + **Reference**:
       - Dutta, S., & Bandyopadhyay, S. (2020). "Characteristics of Fake Job Postings: A Data-Driven Analysis." *Journal of Data Science*, 8(1), 45-58.
       - Anita, M., et al. (2021). "Impact of Job Attributes on Detection of Fake Job Postings." *International Journal of Artificial Intelligence and Applications*, 7(3), 89-102.

These insights and references provide clear explanations of why each research question is important for detecting fraudulent job postings and cite relevant studies that support these findings. They offer a foundation for developing effective machine learning models aimed at enhancing job seeker safety in online recruitment platforms.

# Data Sourcing

## Dataset Details:

The primary dataset, EMSCAD, contains 17,880 job postings from Kaggle. Each job posting is labeled as either fraudulent (1) or non-fraudulent (0). The dataset includes a variety of features to help identify the nature of the job postings, including:

* **job\_id:** Unique identifier for each job posting.
* **title:** The title of the job ad entry.
* **location:** Geographical location of the job ad.
* **department:** Corporate department (e.g., sales).
* **salary\_range:** Indicative salary range (e.g., $50,000-$60,000).
* **company\_profile:** A brief description of the company.
* **description:** Detailed description of the job ad.
* **requirements:** Listed requirements for the job opening.
* **benefits:** Listed benefits offered by the employer.
* **telecommuting:** Indicates if the position allows telecommuting (True/False).
* **has\_company\_logo:** Indicates if the job posting includes a company logo (True/False).
* **has\_questions:** Indicates if the job posting includes screening questions (True/False).
* **employment\_type:** Type of employment (e.g., Full-time, Part-time, Contract).
* **required\_experience:** Level of required experience (e.g., Executive, Entry level, Intern).
* **required\_education:** Level of required education (e.g., Doctorate, Master’s Degree, Bachelor).
* **industry:** Industry of the job (e.g., Automotive, IT, Health care, Real estate).
* **function:** Job function (e.g., Consulting, Engineering, Research, Sales).
* **fraudulent:** Classification attribute where 1 indicates a fraudulent job posting and 0 indicates a non-fraudulent job posting.

Data Cleaning and Preprocessing

The dataset will be cleaned further to handle any remaining missing values and outliers in preparation for modeling. Text preprocessing techniques, such as removing stopwords and stemming, will be applied to enhance the quality of textual data. Exploratory Data Analysis (EDA) will be conducted to gain insights into the dataset's distributions and correlations.

## Exploratory Data Analysis (EDA)

Further exploration will be performed to understand the relationships between variables, such as the distribution of job postings by country, industry, and employment type. Visualizations like bar charts and heatmaps will be used to highlight trends and patterns in the data.

# Analytic Approach

## Target Variable:

The target variable is the 'fraudulent' label indicating whether a job posting is fake (1) or real (0).

## Problem Type:

This is a binary classification problem where the goal is to classify job postings as either fake or real.

## Models to be Evaluated:

I will evaluate the following models:

* **Logistic Regression:** A simple yet effective linear model for binary classification.
* **Random Forest Classifier:** An ensemble method that improves accuracy by averaging multiple decision trees.
* **Support Vector Machine Classifier:** A powerful model for high-dimensional data classification.
* **K-Nearest Neighbors (KNN):** A non-parametric model that classifies based on the closest training examples in the feature space.

Evaluation Metrics**:**

I will use accuracy, precision, recall, and F1-score to evaluate model performance. These metrics provide a comprehensive view of the models' ability to correctly classify fake and real job postings.

Each evaluation metric serves a specific purpose:

* **Accuracy**: Accuracy shows how often the model correctly predicts if a job posting is fake or real. It gives a general idea of prediction correctness. However, if there are many more real job postings than fake ones, high accuracy might not tell the whole story.
* **Precision**: Precision measures how many of the job postings predicted as fake are actually fake. It's crucial because it indicates how reliable the model is in flagging fake postings. High precision means fewer real jobs get incorrectly flagged as fake, which helps maintain trust in the detection system.
* **Recall**: Recall measures how many of the actual fake job postings the model identifies correctly. It's essential for catching as many fraudulent postings as possible to protect users from scams. High recall means fewer scams go undetected.
* **F1-score**: The F1-score combines precision and recall into a single metric, giving a balanced assessment of the model's performance. It's useful because it considers both false positives (real jobs flagged as fake) and false negatives (fake jobs missed). A high F1-score indicates that the model effectively balances between identifying fraudulent postings and minimizing misclassifications of genuine postings.

# Implementation Plan:

The models will be implemented using Python's scikit-learn library. I will preprocess the data, engineer new features, and split the data into training and testing sets. Each model will be trained and evaluated on the test set. The best-performing model will be selected based on the evaluation metrics.

# Conclusion

In summary, this project aims to build a machine learning system that can spot fake job postings accurately. By using models like Logistic Regression, Random Forest, SVM, and KNN, the system will be helpful for job seekers safe from scams and maintain trust in online job platforms which I wish. The plan is to develop, test, and deploy this detection system carefully to ensure it works effectively.