Case NLP

Introduction

If you are reading this, you are doing a second round of interview for the Data Consultant position at Data Wizards. We are looking forward to seeing your problem solving skills in action!

Context

Our prestigious customer WizardsIn has provided us with a dataset of around 18000 job descriptions. Around 800 of them are fake! Thankfully, you are here to help them combat fraud. Your mission is to build a machine learning model to classify the job postings and detect the fraudulent ones.

Dataset

You are provided with a very simple dataset, with a description of the job posting along with the label.

Here is a small table to clarify the dataset:

Field	Description
description	The description of the job
fraudulent	Whether the job is classified as fraudulent or not. 0 = Non-fraudulent 1 = Fraudulent

The dataset (and modeling task) is rather straightforward as the emphasis is on building a solution that is actionable and valuable. Indeed, we are building **integrated** Data & AI **solutions** at Data Wizards.

Deliverable

We expect the following for this take home:

- Create a machine learning model that performs the binary classification
- Add an API endpoint (using Flask or FastAPI for example) and a Dockerfile to run the application
- A simple README on how to run or reproduce the solution

When you are done, you can share the repository with us and we will review it before the second interview.

You can also foresee a small presentation to illustrate your approach, results and insights on the dataset. Should you have any question, do not hesitate to reach out.

Good luck!