Project Proposal

Question/needs:

Employment scams are on the rise. According to CNBC, the number of employment scams doubled in 2018 as compared to 2017. The current market situation has led to high unemployment. Economic stress and the coronavirus's impact have significantly reduced job availability and job loss for many individuals. We need to create a classifier that will have the capability to identify fake and real jobs.

Data description:

The data for this project is available at Kaggle — https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction. The dataset consists of 17,880 observations and 18 features.

A brief definition of the columns is given below:

#Cl : it's mean Colum

Cl n	Cl title	Cl Description
1	job_id	Identification number given to each job posting
2	title	A name that describes the position or job
3	location	Information about where the job is located
4	department	Information about the department this job is offered by
5	salary_range	Expected salary range
6	company_profile	Information about the company
7	description	A brief description about the position offered
8	requirements	Pre-requisites to qualify for the job
9	benefits	Benefits provided by the job
10	telecommuting	Is work from home or remote work allowed
11	has_company_logo	Does the job posting have a company logo

12	has_questions	Does the job posting have any questions
13	employment_type	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable
15	required_education	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
16	Industry	The industry the job posting is relevant to
17	Function	The umbrella term to determining a job's functionality
18	Fraudulent	The target variable ◊ 0: Real, 1: Fake

Tools:

The first step to visualize the dataset in this project is to create a correlation matrix to study the numeric data relationship. I will be planning to use pandas and numpy library for visualization the data. also I will use Logistic regression, Decision Tree and Random Forest if I need it. those are the tools I can think of in the begin. However, going through the camp I will come up with model approaches will be help me.