BDC800 - Capstone Project Proposal

Date: 1/29/2023
Team Number: Group 2
Team Member Names: Nafis Ahmed, Sammar Abbas, Yajing Zhou
Title of the project: “Fake Job Posting Detection: ML Study of Real vs Fake Postings”

Summary:

For the capstone project Winter semester 2023, we aimed to build a machine learning model that will primarily help job seekers in identifying real job postings. As in recent time number of fake job positing is increasing and those who are looking for jobs they become victim of the scam and endup in providing their personal information to them. Model will build on clafficiation technique using publicly available dataset on kaggle “https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction”. Owner of the data has provided permission to copy, modify to perform work for any purpose.

Team Information and responsibilities:

Task & Responsibilities	Nafis	Sammar	Yajing
Data Collection		✔️	✔️
Data Cleaning	✔️	✔️	
Feature Scaling and Data Balancing	✔️	✔️	
Model Selection		✔️	✔️
Model Training & Validation	✔️	✔️	✔️
Model Evaluation		✔️	✔️
Model Testing	✔️		✔️
NLP for Description	✔️	✔️	
Report Writing	✔️	✔️	\

Overview of the problem:

We are planning to build a machine learning model that can classify job postings accurately as Real or Fake job postings. We are also attempting to identify key features that are relevant in the process of classifying job postings and determining its authenticity.

As a result, job seekers will be able to save time and be more safe with sharing their personal information to fake job postings. In addition, this will provide enhanced quality to the overall job searching experience.

Description of technical challenges:

While working on the project, technical challenges we will have are:

1- Dataset is unbalanced in ratio of real and fake job description. Out of 18k total posting only 800 are fake posting. This might result into a biased model in any of the one way.

2- Too many features may increase the complexity of the model which will endup into overfitted model.

3- Right use of NLP in understanding description of the posting will be difficult task.

4- Data cleaning and deduplication may be technical challenges

Proposed solution:

To achieve optimal goal classification model will be used that will predict between fraudulent or real job descriptions. Identifying key traits/features of job descriptions which are fraudulent in nature.

Dataset's Description

This dataset was acquired from Kaggle and contains around 18,000 job descriptions where about 800 are fraudulent. The data consists of both textual information and meta-information about the jobs.

Column Name	Description
job_id	unique job ID

title	the title of the job ad entry
location	geographical location of the job ad
department	corporate department
salary_range	indicative salary range
company_profile	a brief company description
description:	the details description of the ad
requirements	enlisted requirements for the job opening
benefits	enlisted offered benefits by the employer
telecommuting	true for telecommuting positions
has_company_logo	true if company logo is present
has_questions	true if screening questions are present
employment_type	full-type, part-time,contract,etc
required_experience	executive,entry level,intern,etc
required_education	doctorate, master's degree, bachelor,etc
industry	automotive, IT, health care, real estate, etc
function	consulting, engineering, research, sales, etc
fraudulent	0 - real 1 - fake

Project Delivery

The delivery of the project will contain a detailed project report that will contain the summary of project and its methodology alongwith findings and conclusion. Report will contain tables and graphs to visualize data to make it easier to understand for users. Source code will be provided to help others if they want to work on it in the future with same dataset or new dataset, so they don't start working from beginning rather they have something to work on.