ALY 6140 Module 2 — Capstone Project Proposal

Student’s name: Mohammad Hossein Movahedi

Assignment title: ALY 6140 Module 2 — Capstone Project Proposal

Course number and title: ALY6140 71379 Analytics Systems Technology SEC 12 Fall 2022 CPS [TOR-B-HY]

Term: 202315_1 Fall 2022 CPS Quarter

Instructor’s name: Jay Qi, Ph.D.

Nov 14, 2022,

Dimension of the Dataset

A dataset with 18K job descriptions, of which about 800 are fake, is the one I selected for this project. The information is made up of both textual and job-related meta-data. Using the dataset, classification models can be built that can identify false job descriptions.

I found this dataset on Kaggle. It has 18 columns that are listed below:

Columns	Description	telecommuting	True for telecommuting positions
job_id	Unique Job ID	hascompanylogo	True if company logo is present
title	The title of the job ad entry	has_questions	True if screening questions are present
location	Geographical location of the job ad	employment_type	Full-type, Part-time, Contract, etc
department	Corporate department (e.g. sales)	required_experience	Executive, Entry level, Intern, etc
salary_range	Indicative salary range (e.g. $50,000-$60,000)	required_education	Doctorate, Master's Degree, Bachelor, etc
company_profile	A brief company description	industry	Automotive, IT, Health care, Real estate, etc
description	The details description of the job ad	function	Consulting, Engineering, Research, Sales, etc
requirements	Enlisted requirements for the job opening	fraudulent	target - Classification attribute
benefits	Enlisted offered benefits by the employer		

Rationale for Dataset

This dataset is prefect of prediction modeling and also has a lot of features with different types also it explorers the nature of fake and real job posting which is very interesting topic to explorer for us as grad students ready to enter the market.

Questions to Investigate

Below is the list of questions that can be answered:

1- Is there a correction between the different dimensions of this dataset and why?

2- can we accurately predict fake jobs?

3- what are the most similar job descriptions?

Models

Based on my experience in other subjects the models listed below give the best outputs for small datasets containing strings

1- naive buyers

2- SVM after deep cleaning data with cook's distance method

3- GLM as a base for evaluating other more complex methods

References

Chauhan, A. (2022). Real OR Fake Jobs. Kaggle.com.
https://www.kaggle.com/datasets/whenamancodes/real-or-fake-jobs