Preliminary Report: Group 20

Introduction:

Real / Fake Job postings: The project includes visualizing the data and building a model to predict if any job posting is real or fake given the details about the posting. Our team ended up using this dataset because nowadays many fraudulent companies post fake jobs to collect candidates' personal data and try to misuse it. It would be a helpful tool if the candidate can make sure if any given posting is genuine or not.

The steps in this project would involve normalising the raw dataset given by the university of the Aegean | Laboratory of Information & Communication Systems Security. We will study the data, clean the data, perform exploratory data analysis and create a predictive logistic regression model that uses data columns as parameters to predict which jobs are fraudulent and which are real.

Data Description:

The dataset consists of 17,880 observations and 18 features. The data is a combination of integer, binary and textual datatypes. A brief definition of the main attributes is given below:

Title – Description of the job position.
Location - Geographical location of the job
Department – Information about the department of the offered job
Company profile - A brief description of the company.
Requirements - Pre-requisites to qualify for the job
Telecommuting - work from home or remote work allowed or not in the job.
Benefits – Enlisted benefits provided by the job
Required experience – It can be Executive, Entry level, Intern, etc.
Has company logo - Does the job posting have a company logo
Required education – It can be Doctorate, Master's Degree, Bachelor, etc.
Industry - The industry the job posting is relevant to (Automotive, IT, Health care etc.)
Function – Information about the Job's functionality **(**Consulting, Engineering, Research, Sales etc.)
Employment Type - Full-type, Part-time, Contract, etc.

Proposed Analysis:

- ➔ Our expectation for this project work is to analyse relevant features from the dataset and draw conclusions from it. Graphs, charts, or bars will be created to visualize data relativity. We will analyse the relationship between different parameters and deduct the final prediction.

Analysis Methods:

- ➔ Data will be ingested into SQLite and then pulled from it into the python notebook for further analysis. Python libraries, matplotlib and seaborn will be used for the exploratory data analysis. A logistic regression model will be built to predict the output in binary form – the job is fake or real. The python libraries Pandas, NumPy and Scikit-learn will be used for model building.

Milestones:

- ➔ The end goal of this project is to find the parameters which play a significant role in explaining if a job posting is fake or real and build a model which can take any new data of a job posting and predict if it is real or fake with the highest accuracy.

References:

https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction?resource=download

https://mkzia.github.io/eas503-notes/sql.html