



LINKEDGUARD: SAFEGUARDING LINKEDIN PRIVACY BY IDENTIFYING AUTHENTIC COMPANIES, AND SCRAPING RESUMES ACCORDINGLY

PRESENTED BY

Team 6 -

Aydin Ayanzadeh

An Tran

Kiran Reddy

Hrudaya Jinna

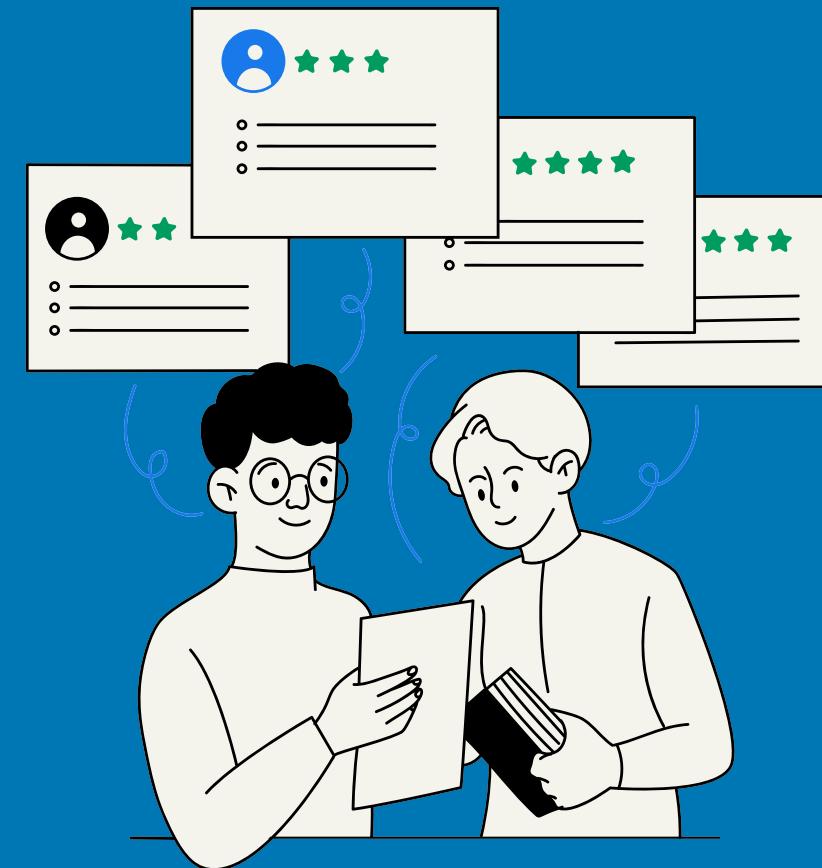




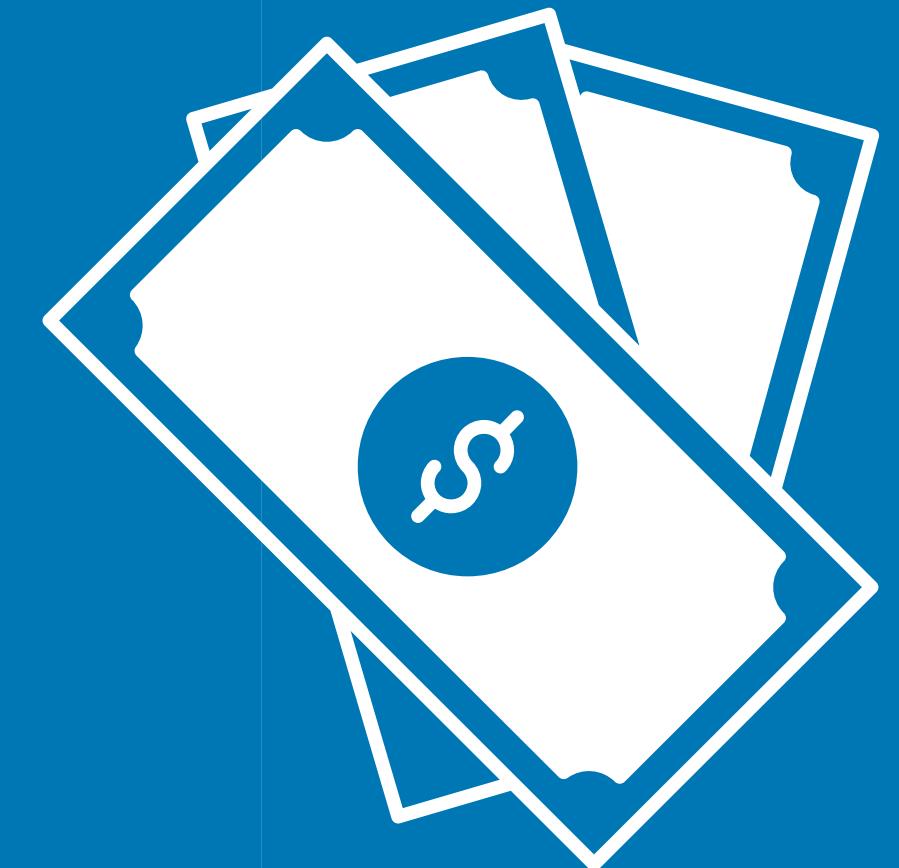
MOTIVATION



Organization



Job posted



Money

A situation where a fake company posted a job listing for a software engineer and requested the applicant to review their products on their website as a task.



DATA BREACH



Most of the time, we provide our resumes to fake companies, and these fake companies steal our information and utilize it to commit fraud.

There is even a data breach instance where LinkedIn has already faced data breaches. One of the fake profile companies obtained information about users from resumes sent to them.



MOTIVATION

In 2019, a fake company called "Asheville Matrix LLC" posted job listings for various positions on LinkedIn. The scammers conducted interviews and even extended offers to unsuspecting candidates. However, the company did not exist, and the scammers were attempting to collect personal information and potentially engage in identity theft.





MOTIVATION

- Email
- Indeed.com
- The way the sentences are framed.
- According to:
<https://www.411.com/phone/1-915-263-8940> ; the phone number is a VOIP (voice over internet protocol), and is "non-fixed". Using VOIP, you can get a new phone number any time you choose; difficult to investigate too.

/pulse/employment-phishing-scam-example-kevin-leo-smith/

Search Home My Network Jobs Messaging Notifications Me For Business

""Good day,

In regard of your resume on INDEED for the Customer Service/Data Entry has been reviewed and you are qualified for a job interview for the position. I want you to setup a Google Hangout and if you have one, add Mr David Becksman the Hiring Manager and IM his on his ID Hangout (davidbecksman@gmail.com) ASAP for the interview and job briefing following our newest online screening introduced, the interview session will commence online.

To Contact our Hiring Manager

Name: David Becksman

Google hangout: davidbecksman@gmail.com

Company Name: Luxottica

Time: ASAP

Best of luck,

Human Resources Department.""--end quote from text message.

OUR APPROACH



Finding out the legitimacy of the company.

Scraping the resume according to risk factor for each company.



LINKEDGUARD



- 1**
Conducted online survey for knowing the factors that affects the legitimacy of the company
- 2**
Based on survey, we selected 11 factors
- 3**
Created dataset using these 11 factors
- 4**
Built ML code for finding out the legitimacy of the company
- 5**
Added risk label ranging from 0 to 5 for each company
- 6**
Predicted risk label
- 7**
Provided this risk label as input to the scrapper
- 8**
Scrapper scraps the resume with respect to risk label

DATASET

Company Name	Company Real or Fake(1 or 0)	LinkedIn Followers > 1000	Profile Picture	Responsive Website	Staff Count > 1000	Staff Premium Accounts	Summary Section	Verified Address	Published Articles	Requests Sensitive Info	Legitimate Email	LinkedIn Recommendations
ALPHALINX INDIA PVT LTD	0	1	1	1	0	0	1	0	0	0	1	0
Buttress Tech0logies Pvt Ltd	0	1	1	1	0	0	1	0	0	0	1	0
Future Focus Tech0logies	0	0	0	1	0	0	0	0	0	0	1	0
Gravity Tech Inc	0	0	0	1	0	0	0	0	0	0	1	0
Digitech Solution Limited	0	1	1	1	1	0	1	0	0	0	1	0
JVS Infotech Pvt Ltd	0	1	1	1	1	0	1	0	0	0	1	0
Cynex Global	0	0	0	1	0	0	0	0	0	0	1	0
FuturesoftTech Infotech	0	0	0	1	1	0	0	0	0	0	1	0
FuturesoftTech point tech0logies	0	0	0	1	0	0	0	0	0	0	1	0
GlobalsoftTech Care Solutions	0	0	0	1	0	0	0	0	0	0	1	0
GlobalsoftTech tech solutions	0	1	1	1	0	0	0	0	0	0	1	0
GlobussoftTech system	0	0	0	1	0	0	0	0	0	0	1	0
GrabsoftTech Tech0logies	0	0	0	1	0	0	0	0	0	0	1	0
HELIOSsoftTech IN0VATIONS	0	0	0	1	0	0	0	0	0	0	1	0
HelixsoftTech Tech0logies	0	0	0	1	0	0	0	0	0	0	1	0
HighsoftTech Tech Lead	0	0	0	0	0	0	0	0	0	0	0	0
HindustansoftTech Software Pvt Ltd	0	0	0	0	0	0	0	0	0	0	0	0
V Solutions	0	1	1	1	0	0	1	0	0	0	1	0
Zoetic Software Inc	0	0	0	1	0	0	0	0	0	0	1	0
Walence Consulting	0	1	1	1	0	0	0	0	0	0	1	0
Skoruz Tech0logies Pvt.Ltd	0	1	1	1	1	0	1	0	0	0	1	0
PIEVIEW TECH0LOGIES	0	1	1	1	0	0	1	0	0	0	1	0
4SPL Tech0logies India Private Ltd	0	1	1	1	0	0	1	0	0	0	1	0

DATASET

Features Name

Legitimate Email

LinkedIn Followers > 1000

Profile picture

Staffs/employees have premium account

Responsive website

Staff premium account

Recommendation

Published Article

verified Address

Request Sensitive Info

Staff count > 1000

**BASED ON QUALITATIVE ANALYSIS
FROM THE SURVEY**

Top-5 Factors

Legitimate Email

Profile picture

Responsive website

verified Address

Request Sensitive Info



DATASET

Risk Label

- All legitimate companies labeled as Risk Level 0
- The fake company is labeled based on a function
- The tuning value is assumed to be 1.19
- n is assumed 5 for top-5 features and assumed 1 for other features

$$R = \begin{cases} 0 & , \quad \text{if company Real} \\ \sum_{i=1}^{11} \alpha^n f_i & , \quad n \text{ is 5 for top - 5 features} \end{cases}$$

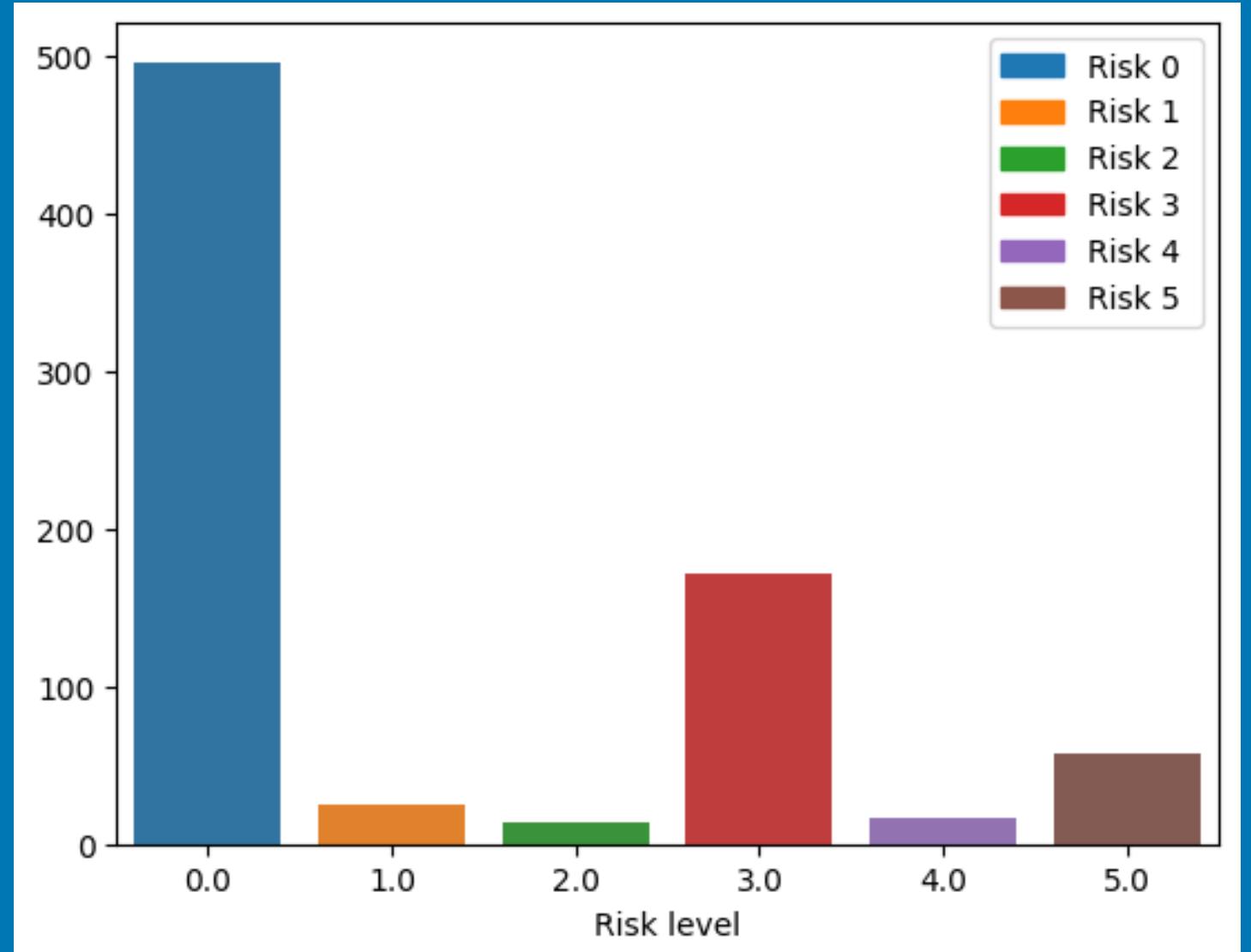
Range	Risk Level
[0-5]	5
(5-7]	4
(7-9]	3
(9-10]	2
(10>)	1



DATASET

Imbalanced dataset

- Synthetic Minority Oversampling Technique (SMOTE)
- Data point frequency increases from 800 to 2976



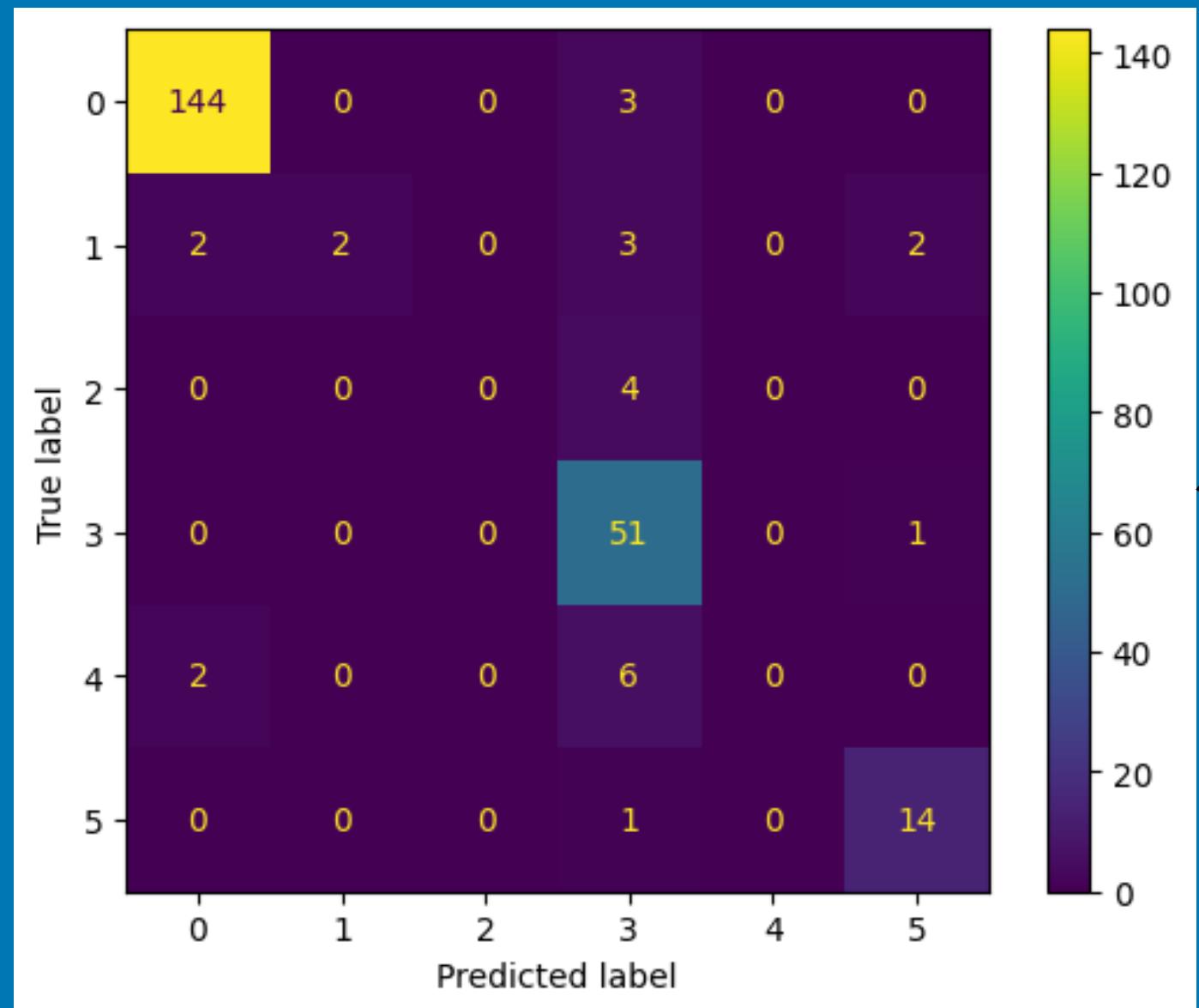
RESULTS

	precision	recall	f1-score	support
class 0	0.97	0.98	0.98	147
class 1	1.00	0.22	0.36	9
class 2	0.00	0.00	0.00	4
class 3	0.75	0.98	0.85	52
class 4	0.00	0.00	0.00	8
class 5	0.82	0.93	0.87	15
accuracy			0.90	235
macro avg	0.59	0.52	0.51	235
weighted avg	0.87	0.90	0.87	235

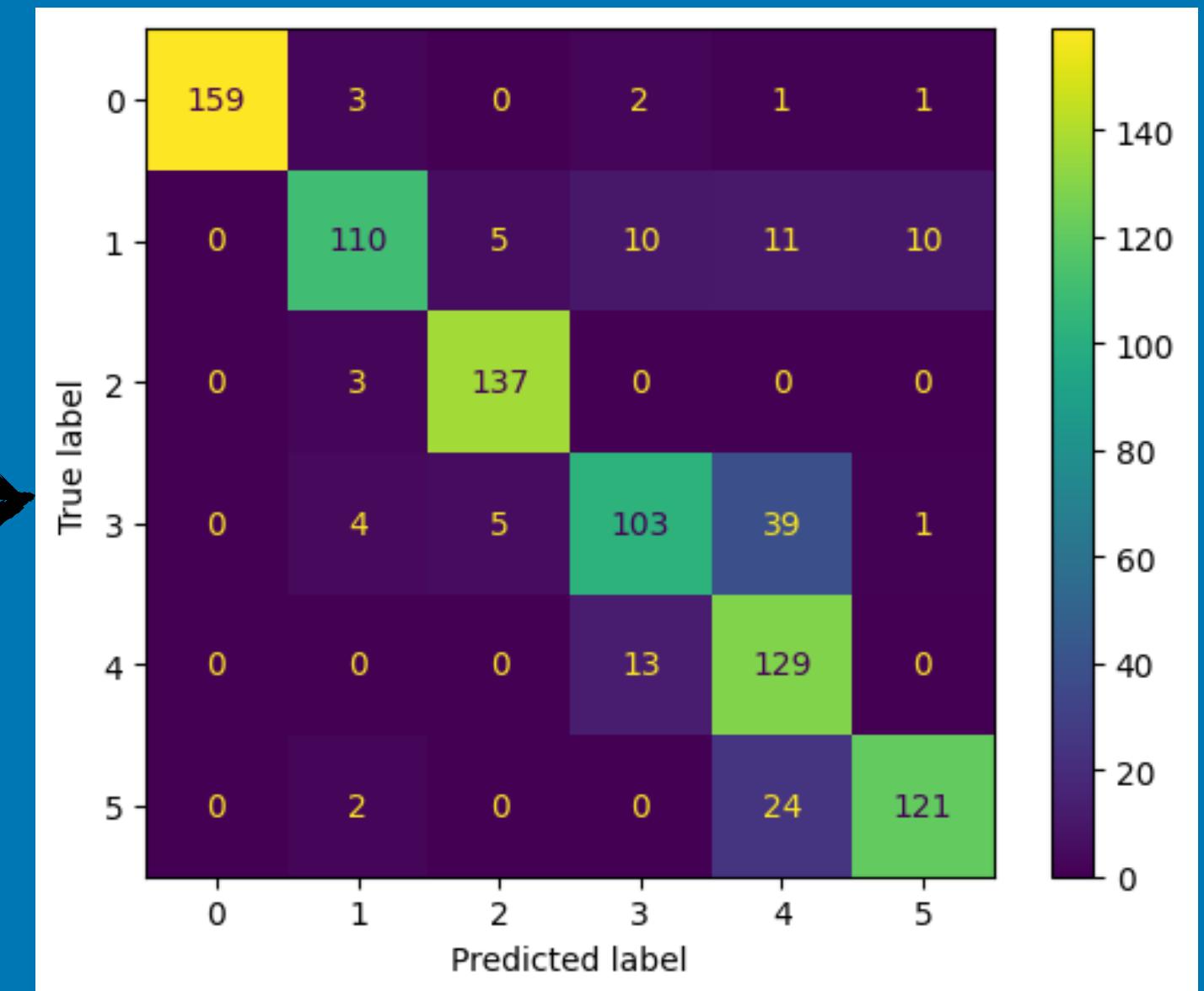
	precision	recall	f1-score	support
class 0	1.00	0.96	0.98	166
class 1	0.90	0.75	0.82	146
class 2	0.93	0.98	0.95	140
class 3	0.80	0.68	0.74	152
class 4	0.63	0.91	0.75	142
class 5	0.91	0.82	0.86	147
accuracy			0.85	893
macro avg	0.86	0.85	0.85	893
weighted avg	0.87	0.85	0.85	893



RESULTS



After over-sampling





RESULTS

Machine learning Algorithm

- DECISION TREE
- SVM
- KNN

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
SVM	0.8724	0.8436	0.8722	0.8403	0.8521
KNN	0.7835	0.9042	0.8218	0.8956	0.8672
Decision Tree	0.9144	0.9126	0.9058	0.9126	0.8941



RESULTS

	precision	recall	f1-score	support
0	0.96	0.93	0.94	143
1	0.96	0.98	0.97	247
accuracy			0.96	390
macro avg	0.96	0.95	0.96	390
weighted avg	0.96	0.96	0.96	390



RISK FACTOR ANALYSIS

Level 0	No Risk	Have a minimal impact. No further action is required
Level 1	Low Risk	Low likelihood of occurring. It can be monitored and managed through routine procedures.
Level 2	Moderate Risk	A moderate impact . Require careful monitoring and management to prevent or mitigate their impact.
Level 3	High Risk	Have a significant impact . Require immediate attention and proactive management to prevent or mitigate their impact.
Level 4	Very High Risk	Have severe consequences . Require urgent action and special measures to prevent or mitigate their impact.
Level 5	Extreme Risk	Have catastrophic consequences. Require immediate and decisive action to prevent or mitigate their impact.

In real-time scenarios, the risk classification should be reviewed regularly and updated as needed to ensure that the risk management strategy remains effective and relevant on a regular basis

CONTACT INFORMATION

- 1 APPLICANT'S NAME
- 2 PHONE NUMBER
- 3 EMAIL ADDRESS
- 4 HOME ADDRESS



APPLICANT'S NAME :- NO RISK

- Primary identifier
- There can be multiple persons with similar names.

PHONE NUMBER:- VERY HIGH RISK

- Hack the phone when comes to female candidates
- May receive unsolicited or threatening calls from telemarketers, scammers, or other unwanted callers.



EMAIL ADDRESS :-HIGH RISK

- This can hack (Highly sensitive industry like defense or intelligence, high-ranking executive email address)

But it's the good source in order to contact the candidate



HOW MITIGATE THE EFFECTS

- Use a professional email address
- Wait for a response
- Choose passwords strong

HOME ADDRESS :EXTREME RISK

- Including identity theft
- burglary
- Physical harm

Why employers may want to see address on your resume
They expect to see it
They're creating an applicant profile for you
To understand how your location may change the job offer



WHEN YOU CAN'T LEAVE YOUR ADDRESS OFF A RESUME

- If you wish to keep this information private for any reason
- If you have found a job posting on a third-party site

No Address

City, State

Example: St. Louis, MO

City, State and Zip Code

Example: St. Louis, MO 63101

Complete Address

Example: 555 Pine St.

St. Louis, MO 63101

Area or Region

Example: Saint Louis Missouri Region

Relocating

Example: Relocating to Tampa, FL

EDUCATION HISTORY:

- 1 NAME OF THE INSTITUTION**
- 2 DEGREE OBTAINED**
- 3 CGPA**
- 4 GRADUATION DATE**

EDUCATION HISTORY: MODERATE RISK

Schooling to graduate levels, most information is already in the public domain - No risk

If a job seeker's education history includes sensitive information such as transcripts or academic records - A higher risk.

Overall it is a Moderate risk

WORK EXPERIENCE

1 JOB TITLES

2 RESPONSIBILITIES

3 DURATION OF EMPLOYMENT



WORK EXPERIENCE :- MODERATE RISK

Confidential information or trade secrets - higher risk.

Inquire about job seeker's personal details (if someone knows in a company) - moderate risk.

SKILLS : NO RISK

- Showcase their abilities
- Attract relevant opportunities.



PROJECTS: MODERATE RISK

- Part of a school curriculum
- Enhances the capabilities of a person - No risk

If the candidate is experienced there are privacy issues related to the client name and project - High risk



PUBLICATIONS:- NO RISK



- To be shared with a wider audience in the public domain.



ACHIEVEMENTS:- LOW RISK

- Public recognition or awards
- Do not usually reveal any sensitive or personal information
(such as membership in a particular group or organization) -A moderate risk

Overall it's a low risk

PERSONAL INFORMATION :

1 AGE

2 GENDER

3 RACE

4 RELIGION

5 MARITAL STATUS

AGE :- MODERATE RISK

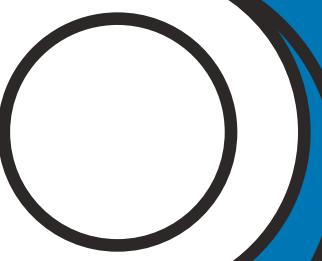
A specific age range ("recent college graduate" or "10+ years of experience") may be necessary or preferred by the employer

On privacy regulations, if one gets the name and age of a person there could be many places with the database having the name and age

GENDER :- MODERATE RISK

- Gender is often readily apparent from the applicant's name
- As per the privacy challenge, this is sensitive information.

RACE:-MODERATE RISK



This can be a highly sensitive issue

When applying for jobs where diversity is a critical factor, in most cases it is not necessary or advisable.

A privacy challenge would classify a race or ethnicity as a Moderate risk

MARITAL STATUS:-MODERATE RISK

Protected information in many countries.

For a position that requires extensive travel or relocation, the employer may be interested in the candidate's marital status.

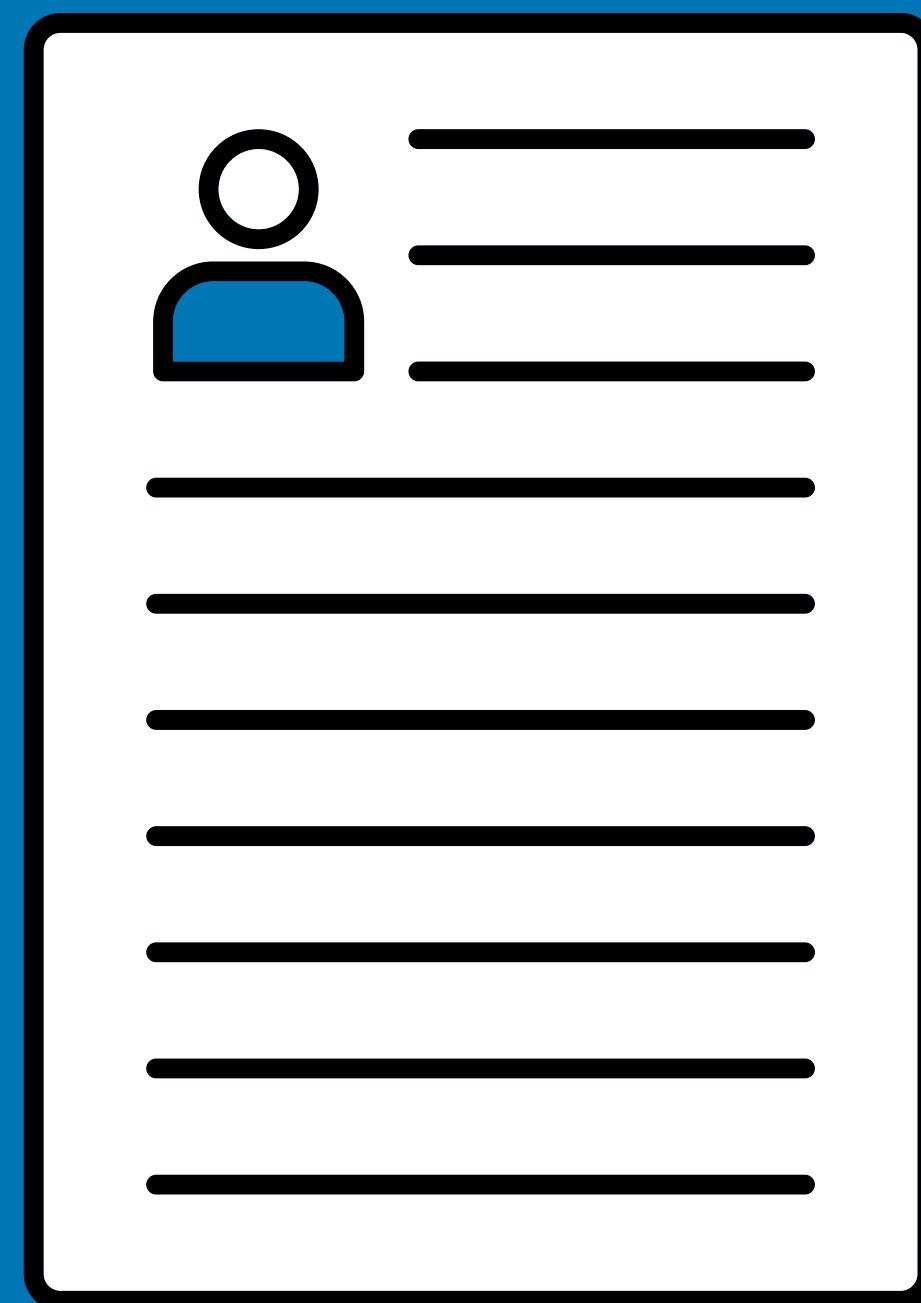
In some cultures or industries, marital status may be considered relevant or important information for identifying a person.

RELIGION :- HIGH RISK ACTIVITY

Religion on a resume can be a highly sensitive issue

A job seeker applying to a religious organization or a faith-based non-profit may choose to highlight their religious affiliation or involvement as a qualification for the job.

Overall personal information has nothing to be done by the company mostly discrimination..But has privacy issues concerned we assign the risk levels



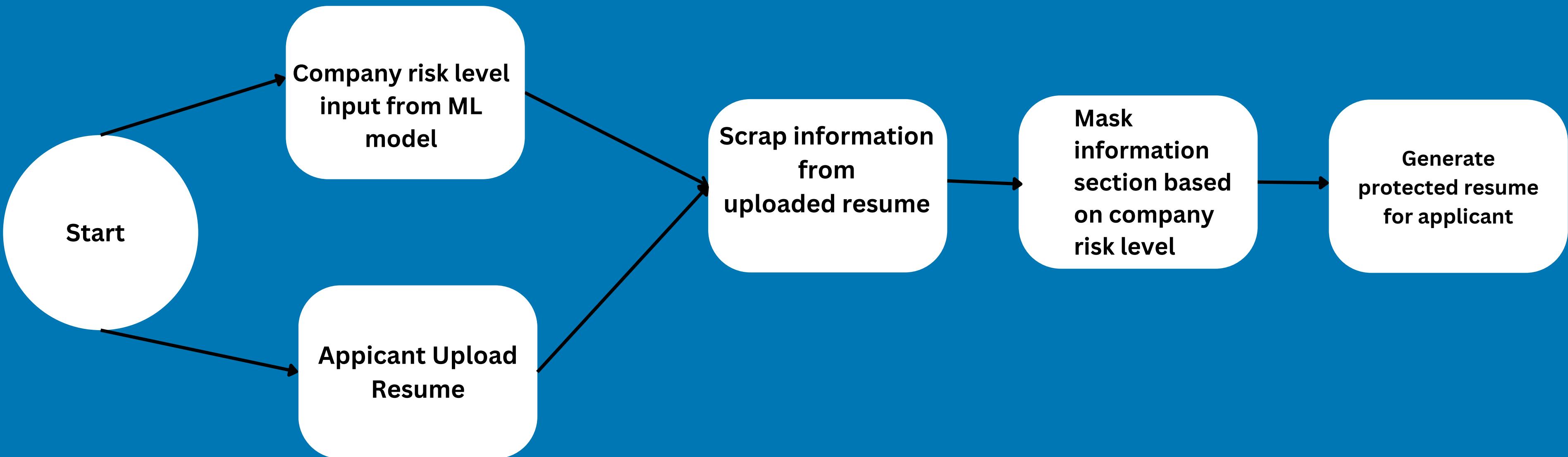
WHAT NEEDS TO BE SCRAPPED

Level 0	No Risk	No changes or scrapping required
Level 1	Low Risk	All features except for those ranked at Risk level 5 (excluding the home address)
Level 2	Moderate Risk	All features except for those ranked at Risk level 5 and Level 4 excluding the home address and phone number)
Level 3	High Risk	All features except for those ranked at Risk Level 5, Level 4, and Level 3 (excluding the home address, phone number, email address, and religion or ethnicity)
Level 4	Very High Risk	All features except for those ranked at Risk Level 5, Level 4, Level 3, and Level 2 (excluding the home address, phone number, religion or ethnicity, education history, work experience, email address, age, gender, race, and marital status)
Level 5	Extreme Risk	All features except for those ranked at Risk Level 5, Level 4, Level 3, Level 2, and Level 1(excluding the home address, phone number, religion or ethnicity, education history, work experience, email address, age, gender, race, marital status)

MAKE SURE..??

- **Research the company**
- **Verify the company's details**
- **Look for red flags**
- **Use reliable job portals**
- **Look at the company profile and the followers**
- **Search online to find information, customer reviews, and complaints about the company**
- **Need to perform a thorough credit check**
- **Gather information about the number of employees it has and other relevant details.**

RESUME PROCESSING WORKFLOW



SCRAPER

Spacy library

Regular
expression

Key words

DEMO



LinkedGuard

FUTURE SCOPE

- Two parts of the project need coordination. We need something like a front end to make these two parts of the project work only from one particular interface.
- When it comes to projects of specific user in LinkedIn, it is given with risk level 2. It means it is of moderate risk. Possibilities of it being in high risk to also be considered in the project.

