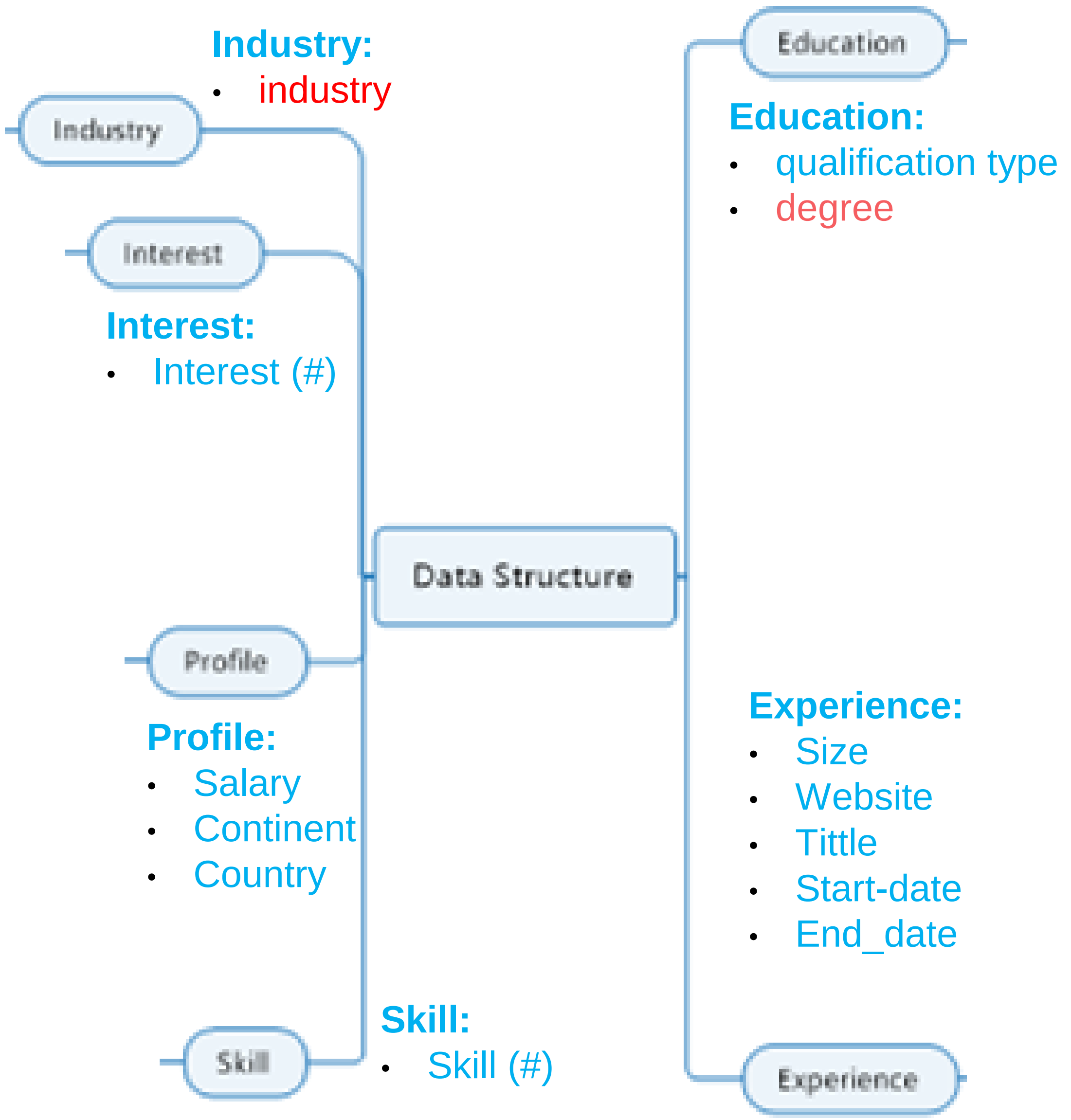
Dataset

In all, we have 1M+ different IDs . The information includes **personal profile, past working experience, education history, industries, interests and skills**. All tables are connected by the column ‘profile_id’ or ‘id’ in some files.

Without building complicated dictionary to category information like **industry** and **degree**. At this stage all the features used by us is shown on the right.

Index Variable

Variable	**Type**	**Definition**
profile_id	str	index of profiles (individuals)



Industry:
- industry

Education:
- qualification type
- degree

Interest:
- Interest (#)

Profile:
- Salary
- Continent
- Country

Experience:
- Size
- Website
- Tittle
- Start-date
- End_date

Skill:
- Skill (#)

Data Preparation

Ideal data structure after cleaning

Variable	Type	Definition
Total tenure	Numeric	Total working tenure except last job
Avg tenure	Numeric	Average working tenure
Other variable	Categorical	Salary level, title, degree… etc.
``		
Target	Binary	If last working experience (if ended) exceeds 1 year

Most valuable information :

1. Previous working experience of each ID
2. Profile info

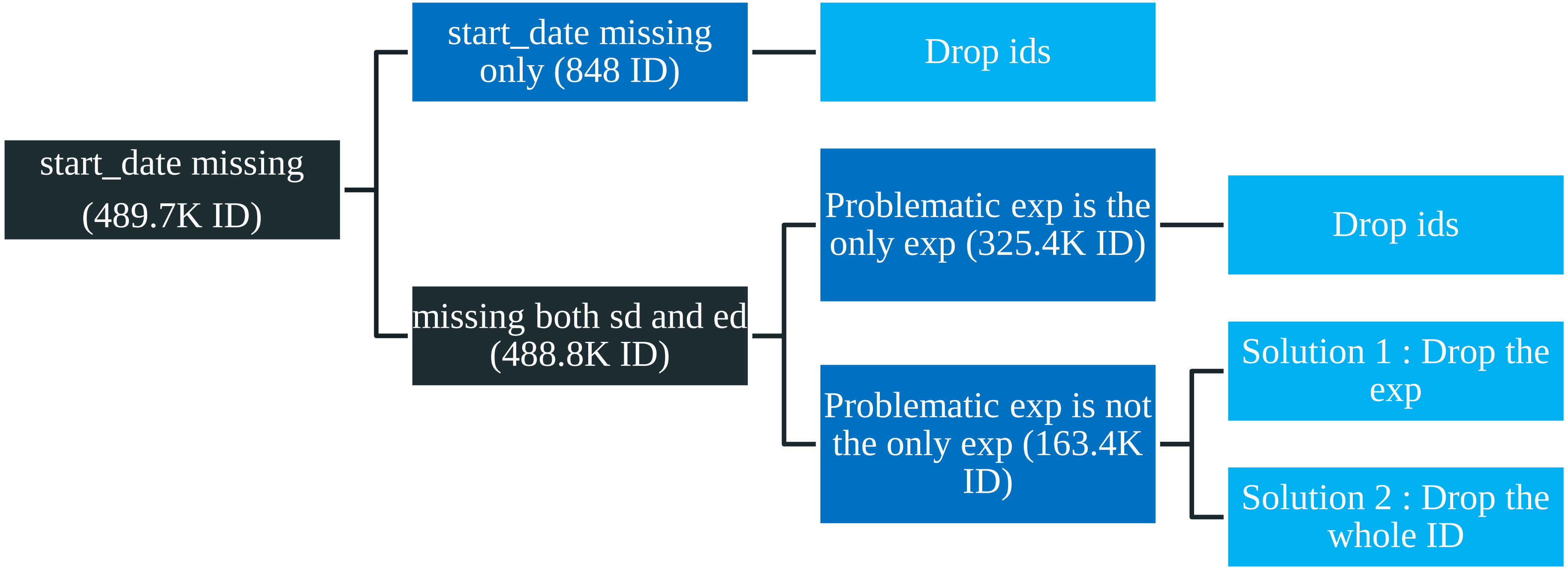
Major Challenges:

1. Lots of missing values in start_date and end_date

2. Unperfect data structure (multiple records with same company, allowing null value in critical variables, etc.)

Date missing problem

- Need one column of dates to be fully notna for determining the sequence of exp
- Want to apply fillna to end_date with start_date later so deal with start_date first. In total we have 1M ids



start_date missing (489.7K ID)

start_date missing only (848 ID)

Drop ids

missing both sd and ed (488.8K ID)

Problematic exp is the only exp (325.4K ID)

Drop ids

Problematic exp is not the only exp (163.4K ID)

Solution 1 : Drop the exp

Solution 2 : Drop the whole ID

After looking into some IDs, the solution 1 seems more reasonable since it keeps more data and doesn't create much bias.

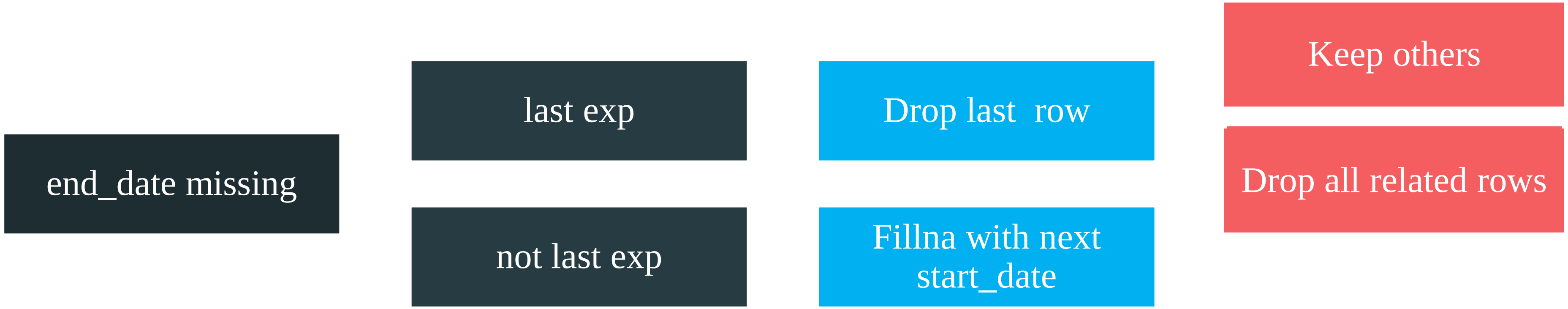
Date missing problem

- Next, deal with end_date missing problem.

- Most exp doesn’t has a end_date because it is the last working exp and had not ended. Hto give up these rows.

- If someone has multi rows for a job in the same company (different positions most likely) and the last row has empty end_date

  - ✓ If tenure > 1 year, keep the rest rows and label that id as ‘loyal’ in the target variable.
  - ✓ If not, drop all info related to that company



end_date missing

last exp

not last exp

Drop last row

Fillna with next start_date

Keep others

Drop all related rows

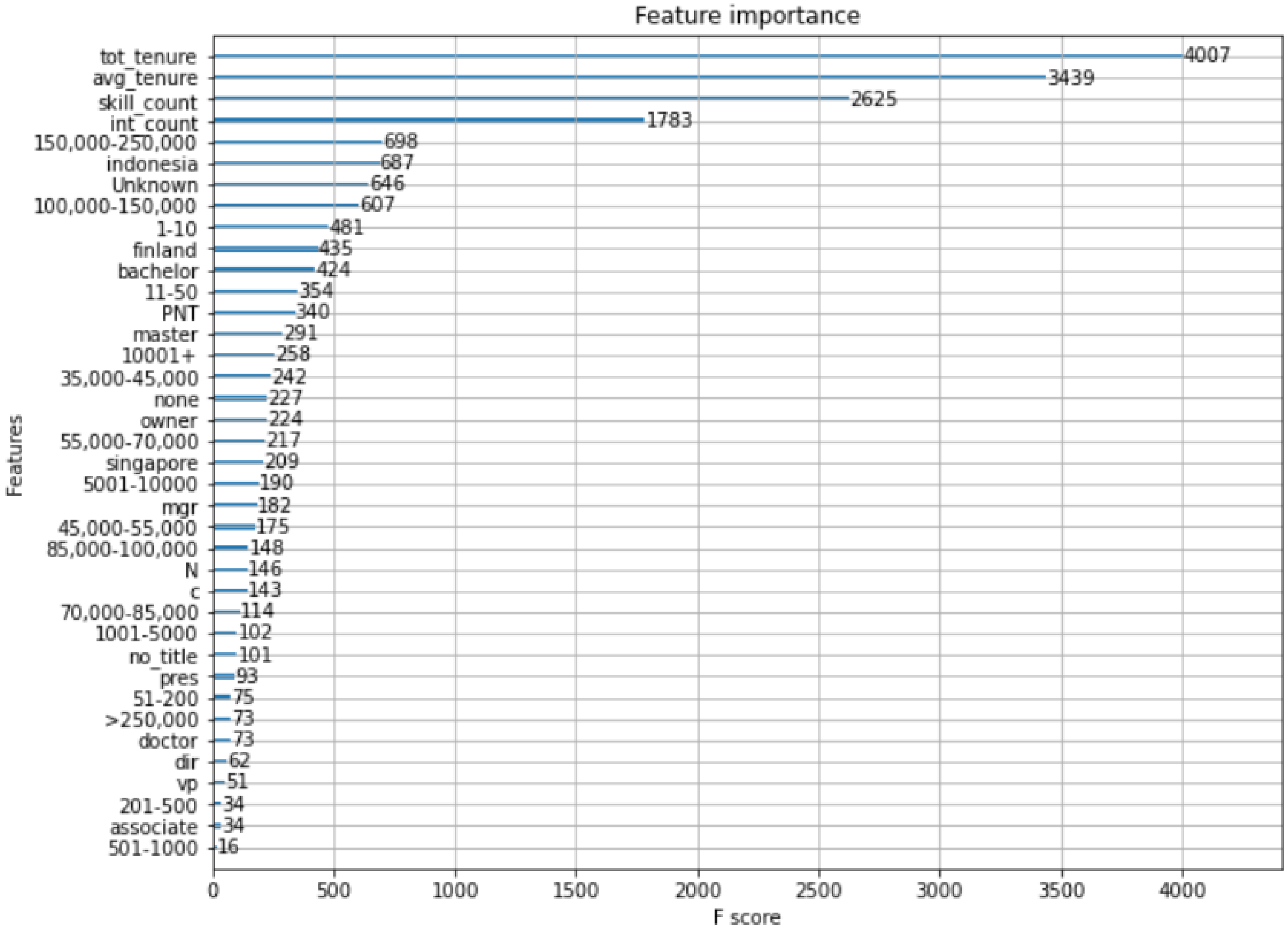
Other exp without end_date can’t be the last one. Fillna with next exp’s start_date

Multi rows of same experience

- The problem is caused by people adding rows when they have a different position even in a same company (which in our case means same 'experience')
- Firstly, drop all IDs that have null value in column 'company'
- Secondly, make some necessary assumptions:
  - Assumption 1: For any ID, same value in 'company' column implies same experience
  - Assumption 2: The total tenure of an exp is determined by the earliest start_date and the latest end_date
  - Assumption 3: If one row has exactly the same start_date & end_date, drop it but keep the ID
- Finally, create a new table in which one row records one unique workid (company + profile_id) and related information.

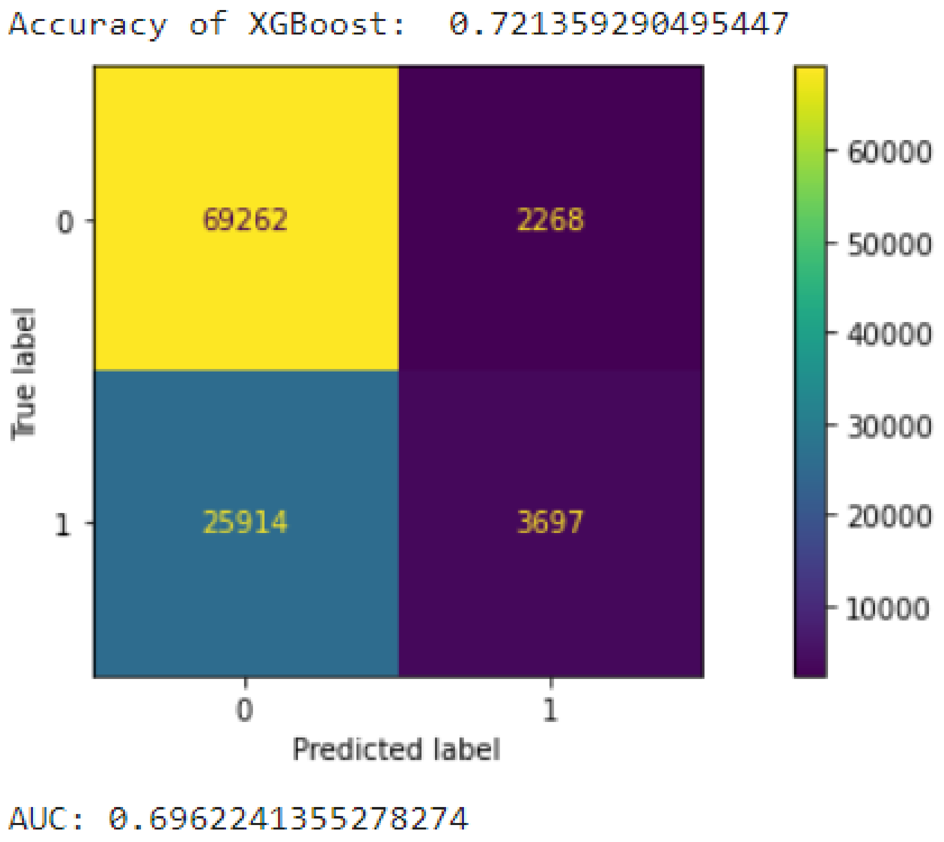
Training data

Variable	Type	Definition
total tenure	Numeric	Total working tenure except last job
avg tenure	Numeric	Average working tenure
size	Categorical	Size of the company in target exp. From 1-10 to 10K+
have_web	Binary	If target exp company has a website
title	Categorical	The title before target exp. From mgr to c and owner
salary	Categorical	The salary level claimed in profile. Over 50% unknow
country	Categorical	Singapore, Finland & Indonesia
continent	Categorical	Asia & EU
int_count	Numeric	How many rows in interest table
skill_count	Numeric	How many rows in skill table
qual_type	Numeric	Associate, bachelor, master and doctor
target	Binary	If last working experience (if ended) exceeds 1 year


Feature importance
tot_tenure 4007
avg_tenure 3439
skill_count 2625
int_count 1783
150,000-250,000 698
indonesia 687
Unknown 646
100,000-150,000 607
1-10 481
finland 435
bachelor 424
11-50 354
PNT 340
master 291
10001+ 258
35,000-45,000 242
none 227
owner 224
55,000-70,000 217
singapore 209
5001-10000 190
mgr 182
45,000-55,000 175
85,000-100,000 148
N 146
c 143
70,000-85,000 114
1001-5000 102
no_title 101
pres 93
51-200 75
>250,000 73
doctor 73
dir 62
vp 51
201-500 34
associate 34
501-1000 16
Features
0
500
1000
1500
2000
2500
3000
3500
4000
F score

XGBoost without much tunning

nthread=-1, seed=42, learning_rate=0.01, subsample=0.5, max_depth=8



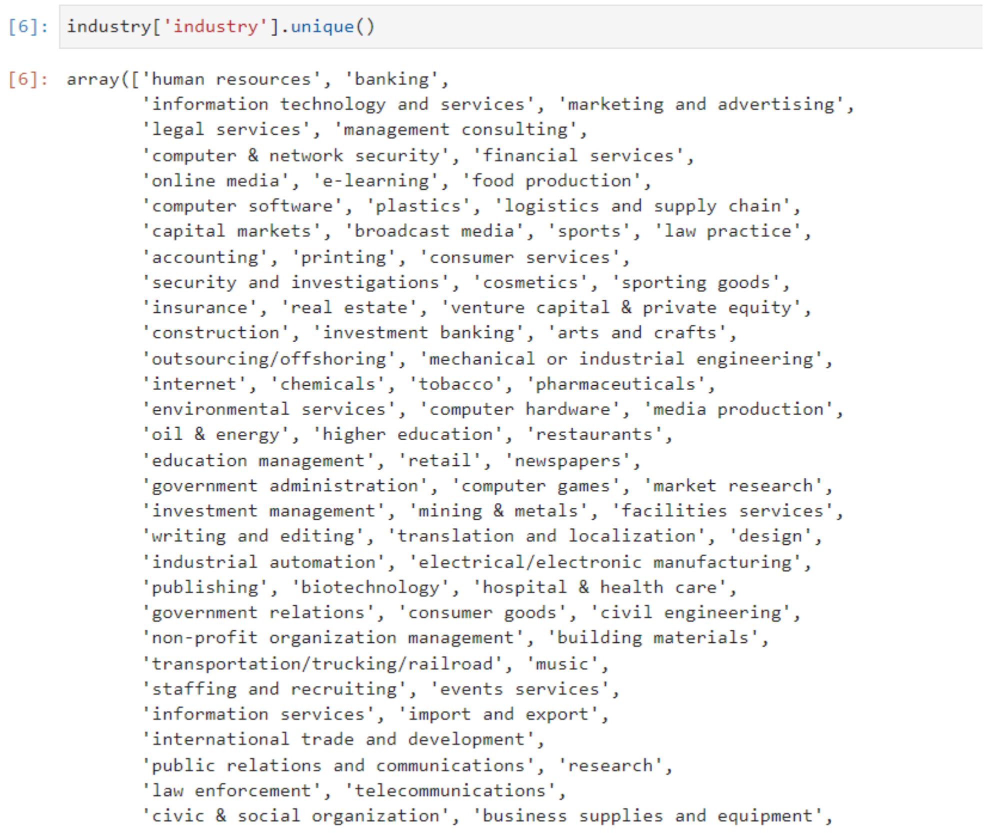
```
Accuracy of XGBoost:  0.721359290495447
```

```
AUC: 0.6962241355278274
```

How to improve the performance

- More features:
  - Require building dictionary, for example ‘industry’ column has 144 unique values, ‘hq_country’ has more than 100 too.
- Further tunning:
  - Maximize the advantages of models like XGBoost and tune carefully with CV
- **Problems to be fixed?**
  - Any dangerous assumptions / missed thought?
  - Other suggestions?

- Throw away data
- LR and tunning



```
[6]: industry['industry'].unique()

[6]: array(['human resources', 'banking',
           'information technology and services', 'marketing and advertising',
           'legal services', 'management consulting',
           'computer & network security', 'financial services',
           'online media', 'e-learning', 'food production',
           'computer software', 'plastics', 'logistics and supply chain',
           'capital markets', 'broadcast media', 'sports', 'law practice',
           'accounting', 'printing', 'consumer services',
           'security and investigations', 'cosmetics', 'sporting goods',
           'insurance', 'real estate', 'venture capital & private equity',
           'construction', 'investment banking', 'arts and crafts',
           'outsourcing/offshoring', 'mechanical or industrial engineering',
           'internet', 'chemicals', 'tobacco', 'pharmaceuticals',
           'environmental services', 'computer hardware', 'media production',
           'oil & energy', 'higher education', 'restaurants',
           'education management', 'retail', 'newspapers',
           'government administration', 'computer games', 'market research',
           'investment management', 'mining & metals', 'facilities services',
           'writing and editing', 'translation and localization', 'design',
           'industrial automation', 'electrical/electronic manufacturing',
           'publishing', 'biotechnology', 'hospital & health care',
           'government relations', 'consumer goods', 'civil engineering',
           'non-profit organization management', 'building materials',
           'transportation/trucking/railroad', 'music',
           'staffing and recruiting', 'events services',
           'information services', 'import and export',
           'international trade and development',
           'public relations and communications', 'research',
           'law enforcement', 'telecommunications',
           'civic & social organization', 'business supplies and equipment',
```