Is this job for real?

Final presentation for MSiA 423

Northwestern

Date: 2020.06.08
Group: 9
Presented by: Joe Zhang
QA: Zach Zhu

Problem: Fake job posting is dangerous and hard to find

Problem

- The scammer uses the job listing to get personal information
- The information is then used to access their bank account or their credit cards and to steal their identity
- The fake jobs also cost people large amount of time, which can be otherwise used to apply for real jobs

Motivation

- It is very difficult for students who have just graduated to judge the authenticity of a job posting
- We lack experiences in the industry and are disadvantaged under the information asymmetry

Solution

- Using machine learning algorithm in the backend
- An web app that predict whether a job posting is fake or real
- Can be integrated as job listing filter for job information providers such Glassdoor, Craigslist, or LinkedIn

Data Description

Raw Data Example

title	Head of Content (m/f)
location	DE, BE, Berlin
department	ANDROIDPIT
salary_range	20000-28000
company_profile	Founded in 2009, we are constantly…
description	Schedules and ensure deadlines…
requirements	Comfortable in a dynamic startup…
benefits	Being part of a fast-growing company…
telecommuting	0
has_company_logo	1
has_questions	1
employment_type	Full-time
required_experience	Mid-Senior level
required_education	Master's Degree
industry	Online Media
function	Management
fraudulent	0

01

Data Source

- Dataset form Kaggle: [Real or Fake] Fake JobPosting Prediction
- Link for the dataset

02

Dataset size

- Size: ~ 60M
- Records: 17880
- Columns: 18

03

Features

- Title, location, department, company profile, salary, JD, requirements, employment type, industry, function, etc.

Model Pipeline: From data wrangling to model training

Cleaning & Engineering

- Drop irrelevant columns
- Imputation
- Text feature cleaning
- Tfidf vectorization
- One-hot encoding
- Standardization

Model Selection

- 5 layers full-connect neural network using relu activation
- Gridsearch: optimizer {sgd, adam, adagrad}, batch_size {8, 16}

Model Training & Save

- Best model: optimizer {adam}, batch_size {16}
- Save engineering pipeline and best model to pickle files for online prediction

Model Result: All success criteria are met

Success Criteria

1. Accuracy > 90%
2. Precision > 0.8
3. Recall > 0.5

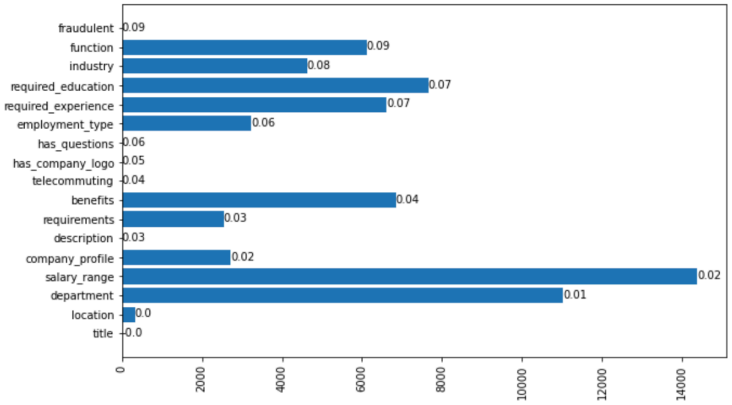
Criteria are met

Test	Predict	
Actual	5060	37
	73	187
Recall	0.719	
Precisioin	0.835	
Accuraccy	0.979	

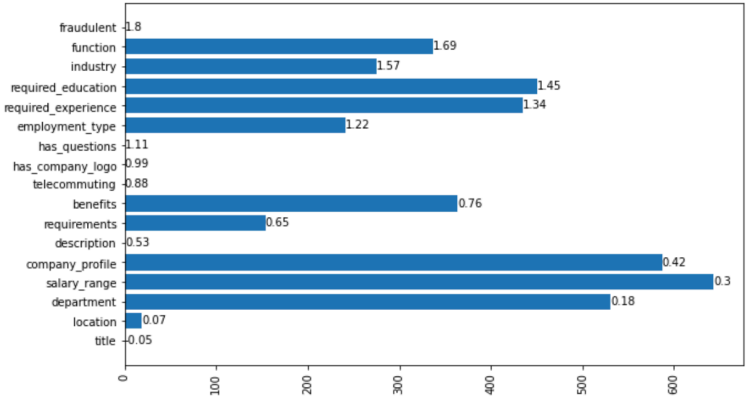
Training	Predict	
Actual	16988	0
	38	827
Recall	0.956	
Precisioin	1	
Accuraccy	0.998	

Interesting Finds (1/3)

Fake job listing are more likely to have NAs in their posting



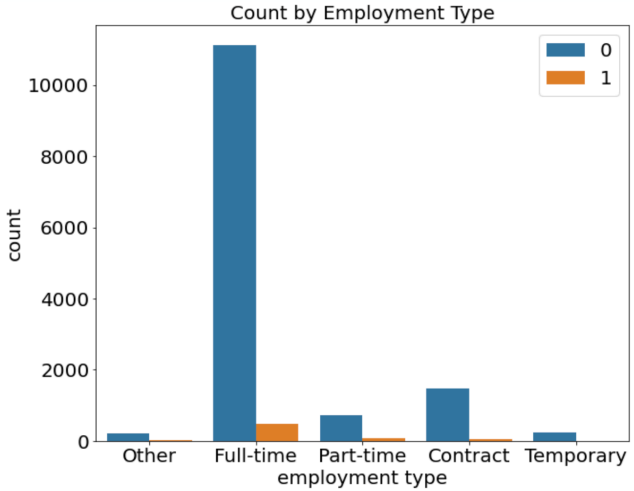
Real

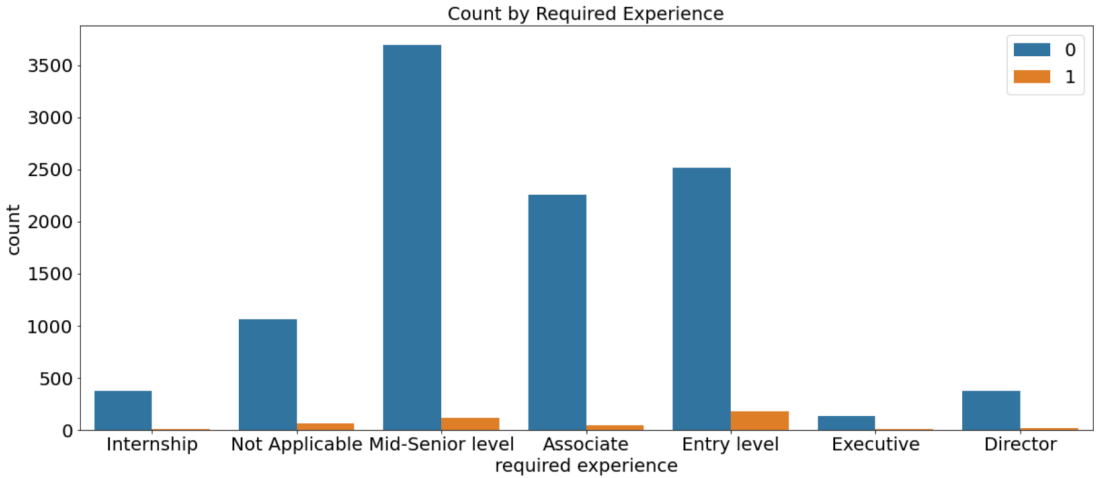


Fake

Interesting Finds (2/3)

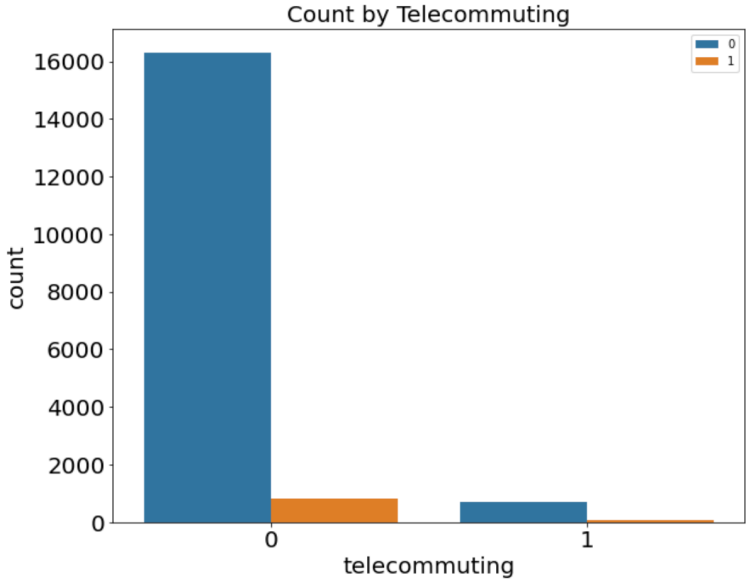
Job postings with requirements of full time, entry level, on-site, and high school degrees are the focus of fraudulent posting

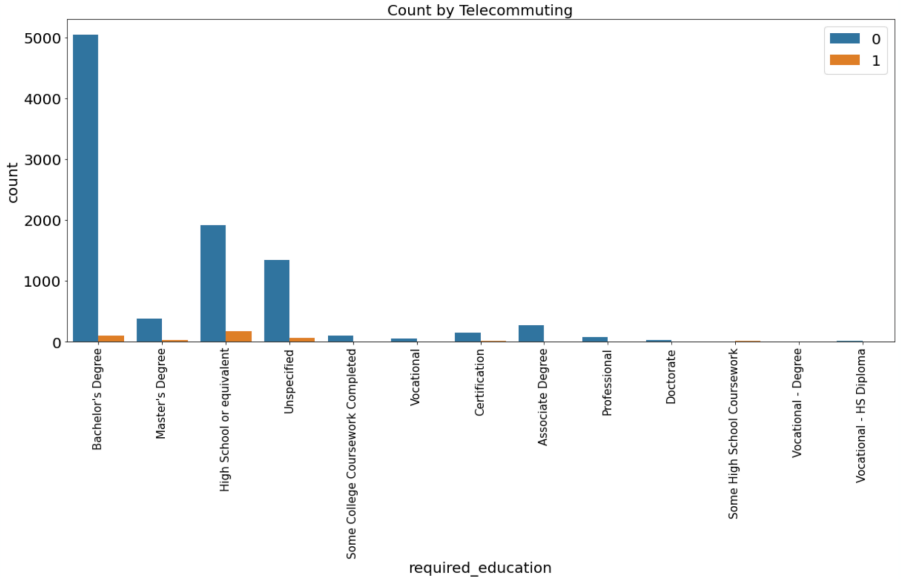




Interesting Finds (3/3)

Job postings with requirements of full time, entry level, on-site, and high school degrees are the focus of fraudulent posting





Conclusion

- Fake job postings are more likely to be related to simple and basic jobs.
- Due to the limited number of fake job postings, it's hard to find very obvious patterns of them. Thus, we have to stay alert when applying for a job and can use machine learning to help us detect the fake postings.
- Hope everybody will get their dream offer in the fall.

Thank You!

Joe Zhang
Master of Science in Analytics
Robert R. McCorick School of Engineering and Applied Science
zihaozhang2020@u.northwestern.edu
Mobile 773.807.5890