# SC1015 Project

## JOB LISTING

By Edward, TianYidong, CuiNan

# Motivation

# Problem Definition

With the recent increases in fraud and scams around the world, and specifically job scams, we will use Natural Language Processing (NLP) techniques to classify job listings as fraudulent or not.

# Data set

Approx. 18k job descriptions, labelled as either fraudulent or not.

# Set the stage

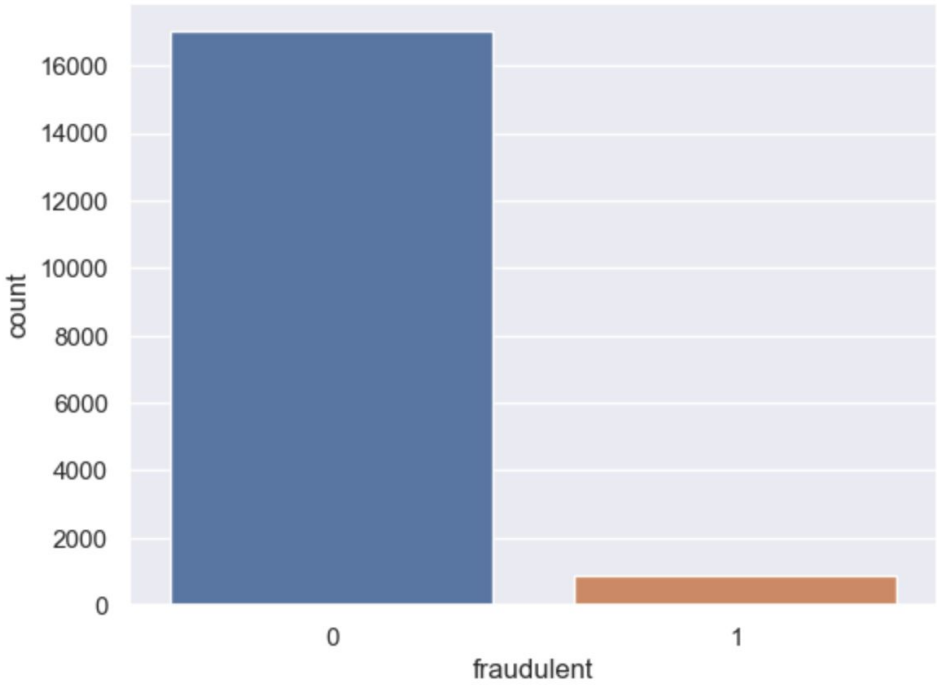
# Exploratory Data Analysis

## Initial data-driven Insights

- Class Distribution
- Word count distribution
- Word frequency

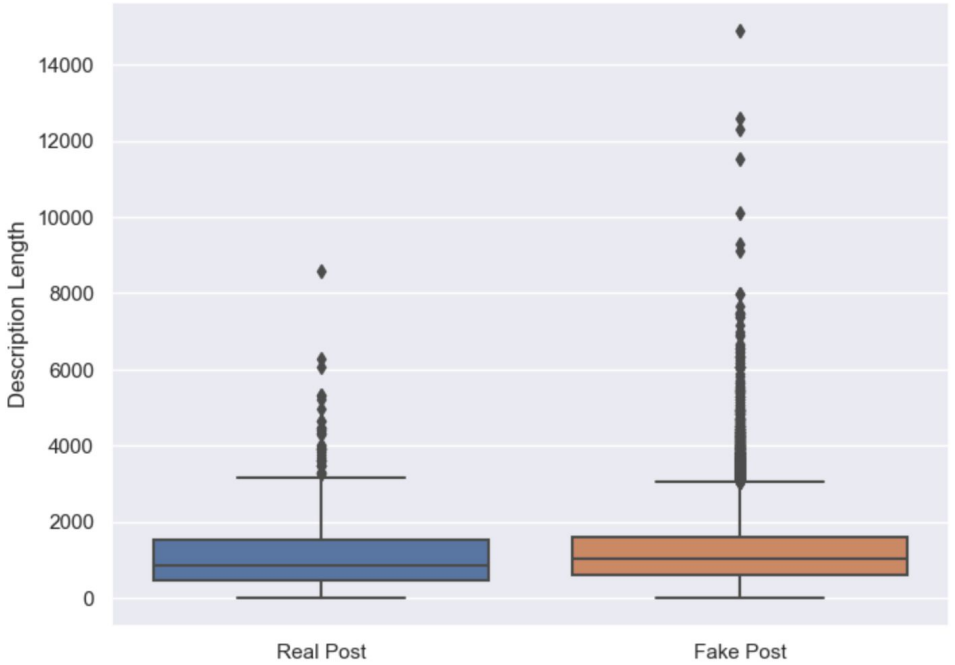
# Class Imbalance

This was later fixed by using some over-sampling techniques such as ADASYN



# Word Count

Fake job listings have many more outliers when it comes to word count



# Overall Plan

1.Spliting data
2.Processing
3.Feature extraction
4.Model selection
5.Model Evaluation
6.Model interpretation

# Data preparation and cleaning

Suit the problem of our choice

Mix of numbers (stored as strings), boolean values and strings

- Data in each row was concatenated, booleans encoded into strings
- NLTK &spaCy
- Post-processing: normalizing case, removing hyperlinks, symbols etc.

# Core analysis

# ML–Classification

- Random forest
- Support Vector Macine (SVM)
- Recurrent Neural Network (RNN)

# Random Forest

- Initially, random forest was chosen as a relatively simple algorithm
- It did not perform well, yielding 0.57 TPR.
- Hence, we decided to use other models.

| Model | Accuracy | F1 | Recall (TPR) | Specificity (TNR) | Precision |
|---|---|---|---|---|---|
| Random Forest (Baseline) | 0.978 | 0.72 | 0.57 | 0.99 | |

# SVM

| Model | Accuracy | F1 | Recall (TPR) | Specificity (TNR) | Precision |
|---|---|---|---|---|---|
| LinearSVC | 0.988 | 0.88 | 0.827 | 0.997 | 0.93 |


- LinearSVC was found to perform better than highr degree kernels
- The SVC was tuned with c-param of 0.1

Confusion matrix of predictions on the test **set**

# RNN

| Model | Accuracy | F1 | Recall (TPR) | Specificity (TNR) | Precision |
|---|---|---|---|---|---|
| Bi-directional LSTM | 0.985 | 0.82 | 0.73 | 0.997 | 0.93 |


- Pretrained GloVe embeddings were used for the input word vector representations
- Bi-directional Long-Short Term Memory (LSTM) was used
- LeakyReLU activation functions were used for the Dense layers, with sigmoid as the output function

Confusion matrix of predictions on the test **set**

# Overall comparison of ML models

| Model | Accuracy | F1 | Recall (TPR) | Specificity (TNR) | Precision |
|---|---|---|---|---|---|
| LinearSVC | 0.988 | 0.88 | 0.827 | 0.997 | 0.93 |
| Bi-directional LSTM | 0.985 | 0.82 | 0.73 | 0.997 | 0.93 |
| Random Forest (Baseline) | 0.978 | 0.72 | 0.57 | 0.99 | |

# Tools and Techniques

Beyond the course

- TFIDF vectorizer
- glove embedding
- ADASYN
- SVM/RNN models

etc..

# Outcome

- Strong correlation between a real job listing and a certain pattern/word use in the job description.
- Possible to classify fake job listings with relatively high accuracy (>80%), despite the highly imbalanced classes.
- Models are able to hit almost 100% accuracy with correctly identifying real jobs.

# Further thoughts

data-driven insights and the recommendations

- Implement more rigorous screening and verification processes
- Improve education and awareness around job fraud
- Identify common patterns and keywords in fraudulent job listings