EE0005 Mini-Project

Classifying Fake Job Postings

Group **Veriton**

Members:
Hoi To
Joel
Monicka
Yashwanth

Outline

Yashwanth

- Objectives
- Exploratory Analysis
- Dataset Cleaning
- Machine Learning
- Conclusion

What is so interesting about this?

- Not easily identifiable
- Very common in known job searching platforms
- Scam of $6.5 M SGD in the the first half of 2021 in Singapore.

Malicious Purpose

Try to steal..

- Personal information
- Money
- Bank details
- Credit card details

What type of ML project?

Classification

- Logistic Regression
- Naive Bayes
- Support Vector Machine
- Random Forest

Our objectives

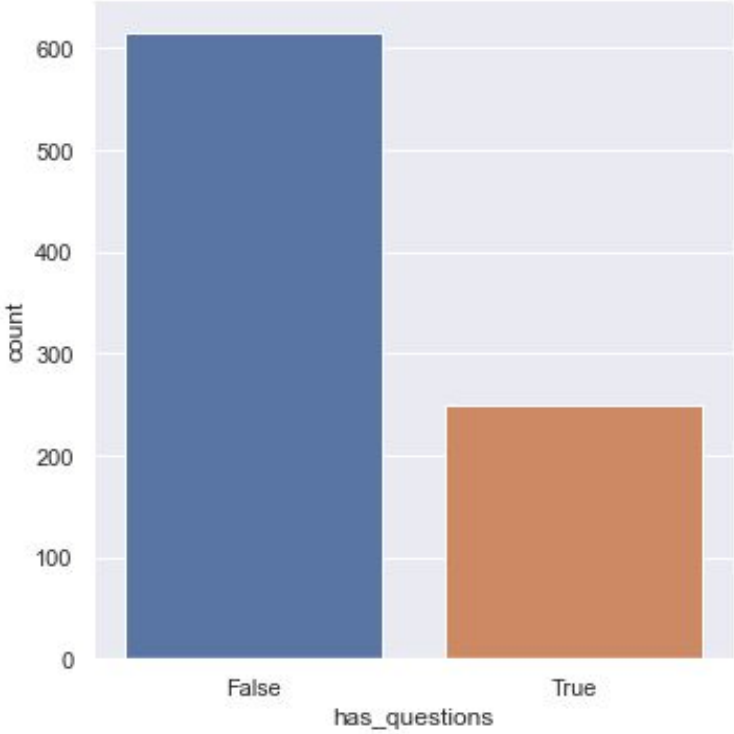
- Finding the common indicators and factors of fake job postings
- Find the relationship between them
- Filter the new job postings
  - implement our model in some job searching sites to filter out fake job postings

Exploratory Analysis

Monicka

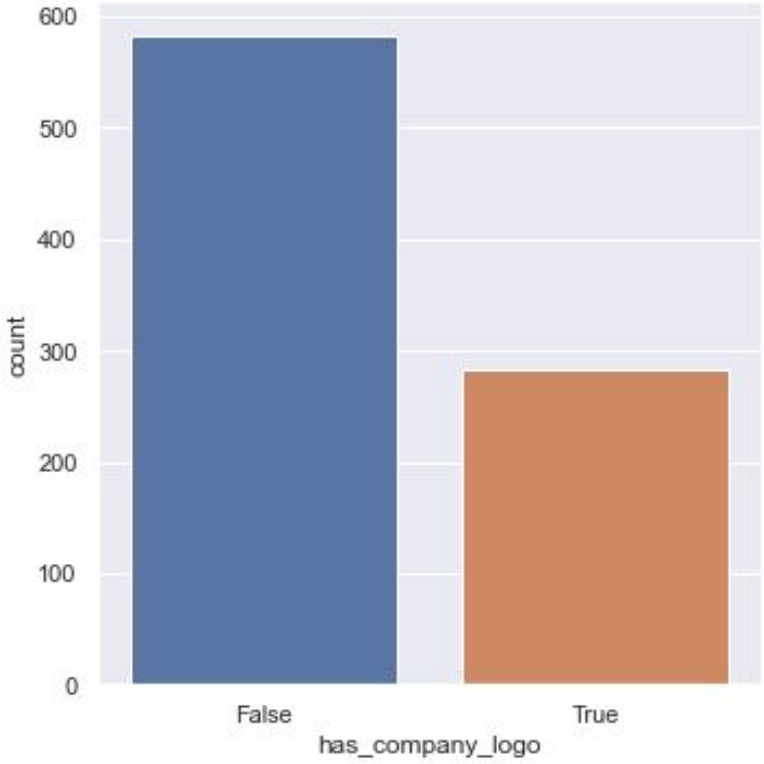
Has_questions variable

- Large portion of fake postings hire *without* an interview
- 71% of fake job postings with False value
- Important determining factor



Has_company_logo variable

- Large portion of fake postings with no company logo
- Only 31% of fake job postings include company logo
- Presence of company logo contributes to company's legitimacy



company_profile	fraudulent
NOCOMPANYPROFILE	587
Aptitude Staffing Solutions has redesigned the recruiting wheel. Our innovative new platform cuts the recruiting time in half, yields scientifically-proven results and clients and candidates enjoy a pleasant experience through advanced, simple to use technology and a tenured, industry-experienced recruiting team. Join us in a fresh new experience of leveraging your career...the way it should be! All represented candidates enjoy the following perks:Expert negotiations, maximizing total compensation package Signing bonus by Aptitude Staffing in addition to client signing bonus (if applicable)1 Year access to AnyPerkRelocation Services for out of town candidatesContinued education in your area of profession, seminars, workshops and other skill development events Contract employees receive quarterly bonuses for the duration of their project Direct-Hire employees receive double bonues ($2,000) per referred/recruited candidate into their newly appointed companyAll candidates are encouraged to participate in our Referral Bonus Program & earn $500 - $1,000 per hired referral	35
Aker Solutions is a global provider of products, systems and services to the oil and gas industry. Our engineering, design and technology bring discoveries into production and maximize recovery from each petroleum field. We employ approximately 28,000 people in about 30 countries. Go to #URL_0fa3f7c5e23a16de16a841e368006cae916884407d90b154dfef3976483a71ae# for more information on our business, people and values.	31
Staffing & Recruiting done right for the Oil & Energy Industry!Represented candidates are automatically granted the following perks: Expert negotiations on your behalf, maximizing your compensation package and implimenting ongoing increases Significant signing bonus by Refined Resources (in addition to any potential signing bonuses our client companies offer)1 Year access to AnyPerk: significant corporate discounts on cell phones, event tickets, house cleaning and everything inbetween. You'll save thousands on daily expenditures Professional Relocation Services for out of town candidates* All candidates are encouraged to participate in our Referral Bonus Program ranging anywhere from $500 - $1,000 for all successfully hired candidates... referred directly to the Refined Resources teamPlease submit referrals via online Referral FormThank you and we look forward to working with you soon! [Click to enlarge Image]	21

Many fraudulent posts come from the same source

Exploratory analysis

Common trends:

- Lack of specific descriptions of the job
- Lower barriers of entry
    - Mostly entry-level jobs
    - Fewer education requirements
    - Less experience required

Dataset cleaning

Joel

To Do:

Handle null values

Convert categorical data into numeric data

Clean text data

Null values

We used the “fillna()” method

```
jobData["function"].fillna(value='Not specified', inplace=True)
jobData['employment_type'].fillna(value='Not specified', inplace=True)
jobData['required_experience'].fillna(value='Not specified', inplace=True)
jobData['required_education'].fillna(value='Not specified', inplace=True)
jobData['industry'].fillna(value='Not specified', inplace=True)
```

One hot encoding

Row	required_experience
1	“Associate”
2	“Director”
3	“Entry level”
4	“Not Applicable”
5	“Entry level”

RE_associate	RE_director	RE_entry_level	RE_not_applicable
1	0	0	0
0	**1**	0	0
0	0	**1**	0
0	0	0	**1**
0	0	**1**	0

Text Cleaning

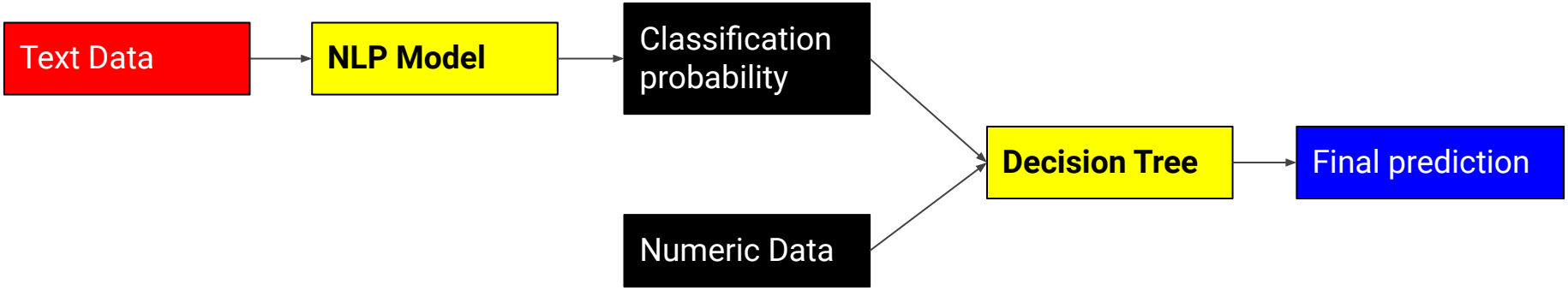
Step 1: filter out all the stop words

Original	Filtered
I like cats, dogs and leisurely strolls	Like cats dogs leisurely strolls

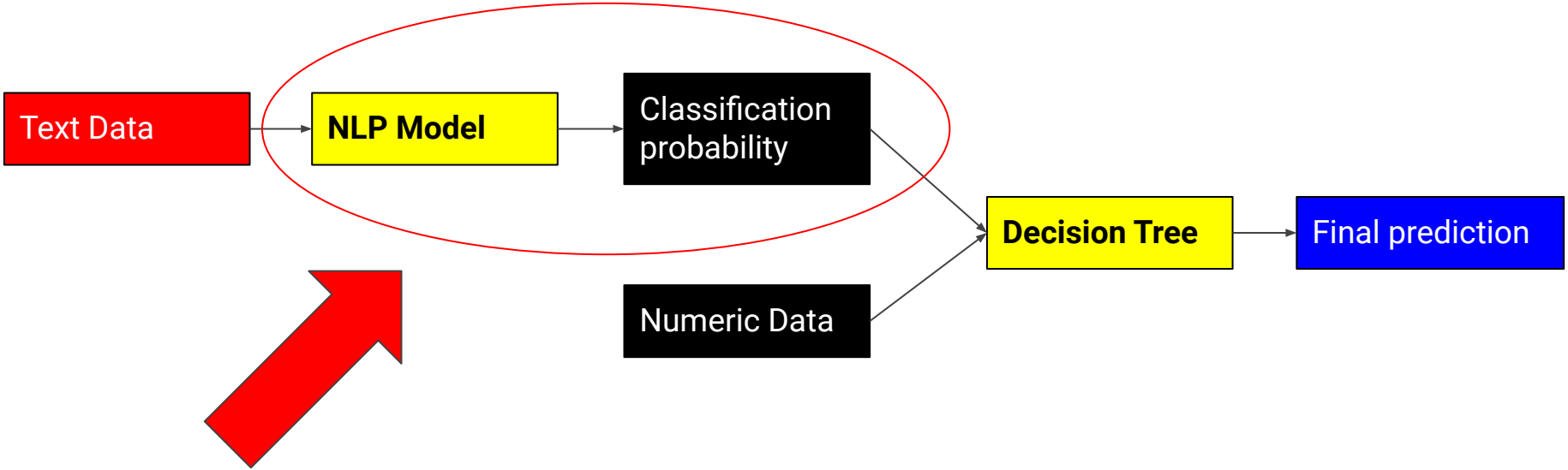
Step 2: reduce all words to their root form, known as stemmatization

Filtered	Filtered & Stemmatized
Like cats dogs leisurely strolls	Like cat dog leisure stroll

Model approach



Model approach



Bag of words

Extracts only word counts of every word, in every data entry

No.	Sentence
1	I like dogs
2	I like running

No.	I	like	dogs	running
1	1	1	1	0
2	1	1	0	1

Naive Bayes model

Problem: What the probability of being fraudulent, given XXX words?

Solution: Ratios, conditional probability & Bayes' theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Implementation

SKlearn provided tools to get word counts

SKlearn also has a Naive Bayes model to train

Finally, we used the model to make predictions

Final outcome: NLP predictor (numeric)

0	-371.901004
1	-778.186145
2	-297.460371
3	-1001.577405
4	-442.564292
	...
17875	-936.704362
17876	-456.605411
17877	-327.425614
17878	-37.266631
17879	-916.090651

Machine Learning

Hoi To

What to do next?

What we have

Raw data

Cleaned data

NLP Predictor

Our objective

To come up with a ML model that can tell whether a job posting is fraudulent

Machine Learning

What ML model are we doing?

Classification models

Only cleaned data + NLP predictor (Numeric values) are fed into the classification models

The following classification models are tried out:

- Decision tree
- Logistic regression
- Support vector machine
- Random forest

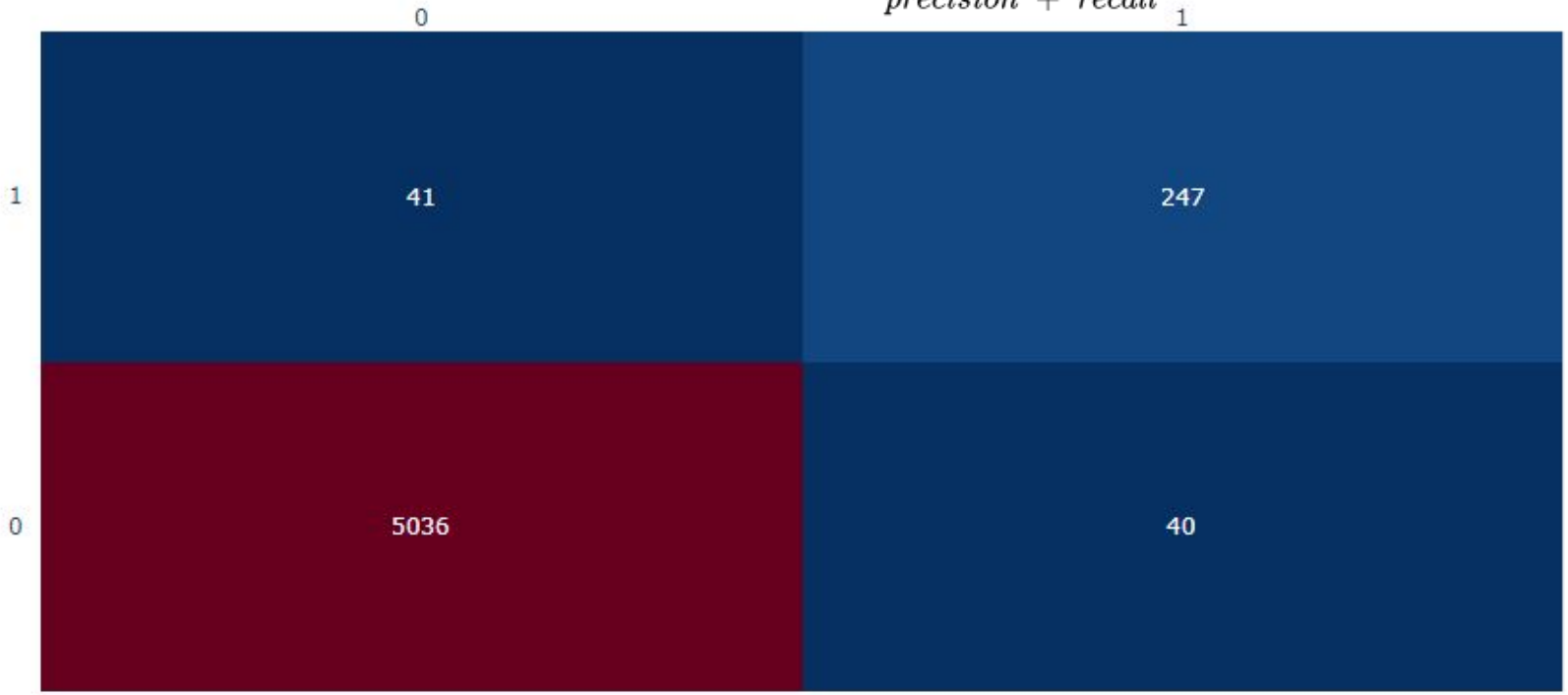
Random forest model so far gives the best result

```
Test data Accuracy:      0.9848993288590604
Test Data Precision:     0.8606271777003485
Test Data Recall:        0.8576388888888888
Test Data F1 score:      0.8591304347826088
```

$$precision = \frac{TP}{TP + FP} = +P$$

$$recall = \frac{TP}{TP + FN}$$

$$F - score = 2\,\frac{precision \times recall}{precision + recall}$$



Random Forest

How it works?

- Random: creates **multiple random datasets** from the raw dataset, select **random features** to train decision trees
- Forest: multiple decision trees
- Make prediction by averaging results of each component tree

Why did we choose it?

- Highest F1-score and accuracy
- Fewest no. of FNs and FPs

Best ML model: Random Forest

Key variables:

In [27]:
```
feature_imp = pd.Series(clf.feature_importances_, index=X_train.columns).sort_values(ascending=False)
feature_imp.head(10)
```

Out[27]:
```
NLP_Pred                              0.476904
has_company_logo                      0.049841
industry_is_Oil & Energy              0.024865
salary_upper_limit                    0.024228
salary_lower_limit                    0.021088
country_code_is_US                    0.020810
has_questions                         0.020546
function_type_is_Administrative       0.015744
required_education_is_Unspecified     0.010592
industry_is_Accounting                0.010429
dtype: float64
```

Conclusion

Monicka

Our mission

- Finding the key factors that constitutes fake job postings
- Finding the best ML model to detect fake job postings

Best model: Random Forest

- Key variables include:
  - Has_company_logo
  - Has_questions
  - Required_education

- In line with variables singled out during exploratory analysis

How is this information useful?

- Climate of job insecurity due to pandemic
- Warn potential job-seekers who are vulnerable
- Implementation of the algorithm to weed out fake postings in job portals

Shortcomings

- Model is specific to dataset
- More work needed to generalise it

Hoi To	Joel	Monicka	Yashwanth
Dataset cleaning, machine learning models (SVM and Decision Tree), presentation and slides, merging and cleaning all the jupyter nb files	Dataset cleaning (on NLP part), exploratory analysis (on text), machine learning models (Decision tree and Logistic regression), presentation and slides	Dataset cleaning, exploratory analysis, machine learning models (Logistic regression), presentation and slides	Dataset cleaning, presentation and slides

Our Task Distributions

Thank you!