Using machine Learning models to detect online fake job postings

IE0005 Mini Project - Team **Koviema**

Our working

Data Visualization and Analysis

Jae Won

Juneyoung

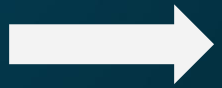
Random Forest Non-NLP Approach

Using NLP to detect fake job postings

Nguyen Tuan Anh

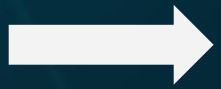
Problem statement

Unemployment rate increased significantly in USA in 2020



Job security was a major issue

Increase of people seeking for job



Increase of fake job posting

Society impact

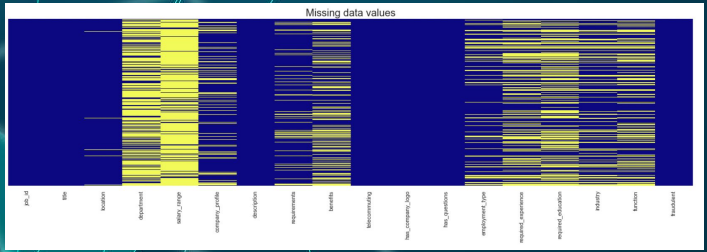
An accurate classification model that can filter out fraudulent data to prevent scammers from having opportunity to scam more desperate people

Data analysis

01

Initial view of dataset
and data analysis

Data cleaning - Country-wise (US)



estions	employment_type	required_experience	required_education	industry	function	fraudulent	country
0	Other	Internship	NaN	NaN	Marketing	0	US
0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	0	NZ
0	NaN	NaN	NaN	NaN	NaN	0	US

```
US_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10656 entries, 0 to 17877
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   job_id               10656 non-null  int64 
 1   title                10656 non-null  object
 2   location             10656 non-null  object
 3   department           3038 non-null   object
 4   salary_range         1557 non-null   object
 5   company_profile      8580 non-null   object
 6   description          10656 non-null  object
 7   requirements         8882 non-null   object
 8   benefits             5974 non-null   object
 9   telecommuting        10656 non-null  int64 
 10  has_company_logo     10656 non-null  int64 
 11  has_questions        10656 non-null  int64 
 12  employment_type      8894 non-null   object
 13  required_experience  6372 non-null   object
 14  required_education   6179 non-null   object
 15  industry             7974 non-null   object
 16  function             6752 non-null   object
 17  fraudulent           10656 non-null  int64 
 18  country              10656 non-null  object
dtypes: int64(5), object(14)
memory usage: 1.6+ MB
```

	job_id	title	location	depar
0	1	Marketing Intern	US, NY, New York	Ma
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	
3	4	Account Executive - Washington DC	US, DC, Washington	
4	5	Bill Review Manager	US, FL, Fort Worth	
5	6	Accounting Clerk	US, MD,	

Identified missing values in data set. Fill in location and create a new column which identifies the country of job posting

Extract job posting located in US and create a new data set

Data cleaning - Data type

employment_type	required_experience	required_education	industry
Other	Internship	NaN	NaN
NaN	NaN	NaN	NaN
Full-time	Mid-Senior level	Bachelor's Degree	Computer Software
Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care
NaN	NaN	NaN	NaN

telecommuting	has_company_logo	has_questions
0	1	0
0	1	0
0	1	0
0	1	1
0	0	0

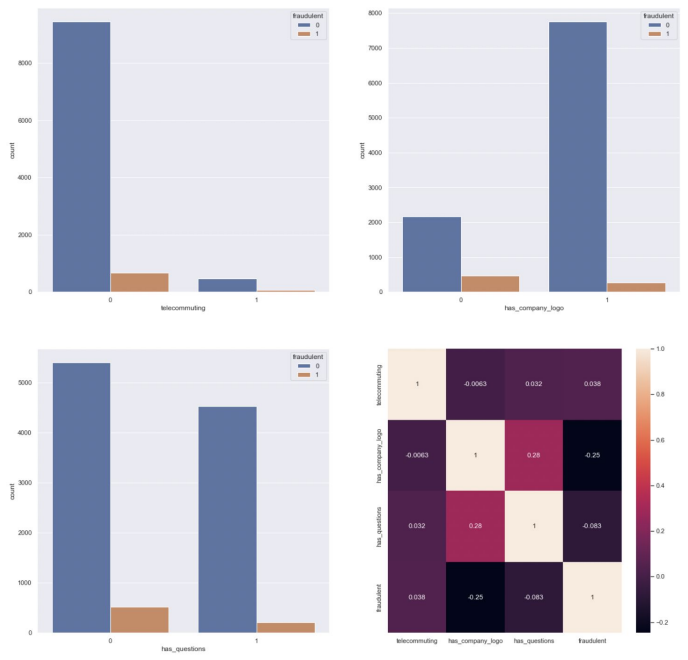
company_profile	description	requirements
We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...
Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...
Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...
SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...
NaN	Job OverviewApex is an environmental consultin...	NaN

Categorical data
(Object)

Numerical data
(Int 64)

Text data
(Object)

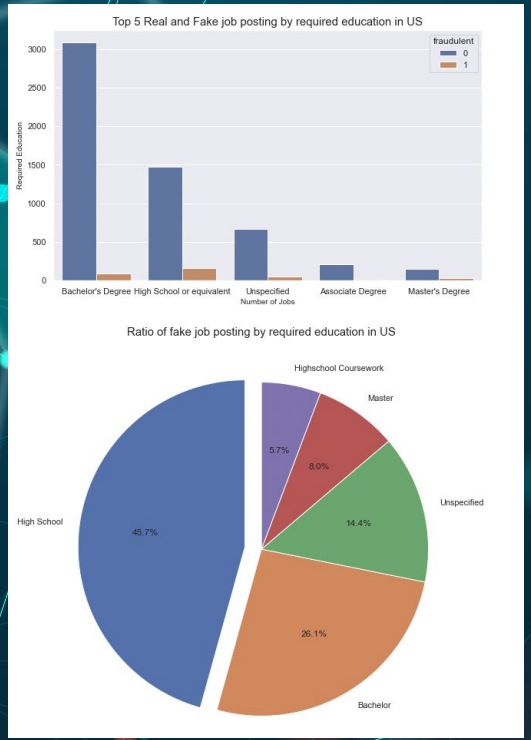
Data analysis - Numerical data



significantly less fake job postings in the dataset. Hard to identify the trend

- identified that fake job postings are less likely to be posted as telecommuting
- Fake job postings mostly have no company logo, unlike real job postings.

Data analysis - Categorical data



Required education

- Total job: Bachelor's Degree
- Fake job: high school

Required experience

- Total job: Mid-senior level
- Fake job: Entry level

Employment types

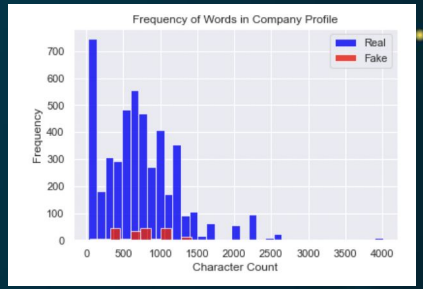
- Total job: Full-time
- Fake job: Full-time

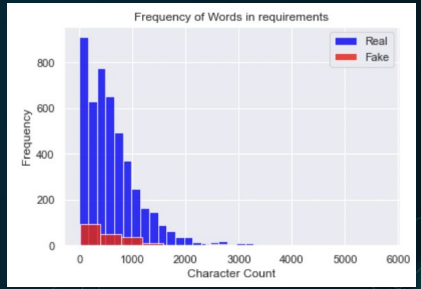
industry

- Total job: Information/technology
- Fake job: Oil & Energy

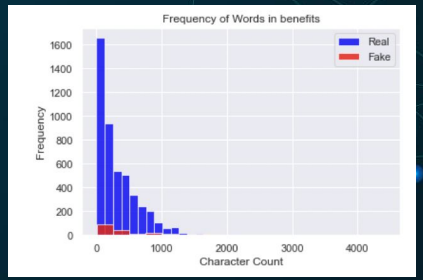
Data analysis - Text data

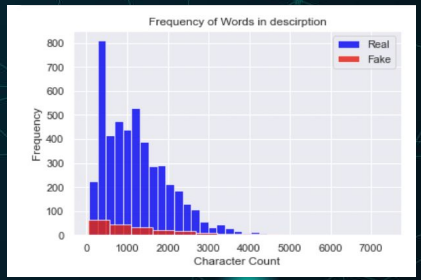
	company_profile	description	requirements	benefits	fraudulent	total_text
3	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	Our passion for improving quality of life thro...
4	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	SpotSource Solutions LLC is a Global Human Cap...
7	Airenvy's mission is to provide lucrative yet ...	Who is Airenvy?Hey there! We are seasoned entr...	Experience with CRM software, live chat, and p...	Competitive Pay. You'll be able to eat steak e...	0	Airenvy's mission is to provide lucrative yet ...
23	WDM Group is an innovative, forward thinking d...	#URL_eda2500ddcedb60957fcd7f5b164e092966f8c4e8...	Job Requirements:A reputation as a "go-getter"...	Businessfriend will offer a competitive six fi...	0	WDM Group is an innovative, forward thinking d...
32	We are an award-winning team of professionals,...	Construction: Entry-Level Craftsman Associate ...	Requirements: (Please do not apply if you do ...	Benefits: Hourly plus commissions. Work with ...	0	We are an award-winning team of professionals,...





Frequency of word significantly low in fake job postings.
Less information provided in fake job postings.





Random Forest

02

Fake job detection using random forest model

Data Exploration - Data Cleaning

US Data (object)

telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function
0	1	0	Other	Internship	NaN	NaN	Marketing
0	1	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service
0	1	0	NaN	NaN	NaN	NaN	NaN
0	1	0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales
0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider

"US" countries extracted from the given Dataset with the data type of "Object" for multiple columns.

US Data (int64)

telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function
0	1	0	4	0	0	NA	0
0	1	0	4	6	0	NA	0
0	1	0	0	2	1	Computer Software	0
0	1	1	0	2	1	Hospital & Health Care	0
0	0	0	4	6	0	NA	0

Columns with "Object" data types has been converted into numerical "int64" data types for comparison.

Data Exploration - Data Cleaning

```
#replacing and cleaning all the values in "required_experience" into numeric values

job_us['required_experience'] = job_us['required_experience'].replace(['Internship'], 0) #Internship = 0
job_us['required_experience'] = job_us['required_experience'].replace(['Entry level'], 1) #Entry level = 1
job_us['required_experience'] = job_us['required_experience'].replace(['Mid-Senior level'], 2) #Mid_Senior level = 2
job_us['required_experience'] = job_us['required_experience'].replace(['Associate'], 3) #Associate = 3
job_us['required_experience'] = job_us['required_experience'].replace(['Executive'],4) #Executive = 4
job_us['required_experience'] = job_us['required_experience'].replace(['Director'],5) #Director = 5
job_us['required_experience'] = job_us['required_experience'].replace(['NA', 'Not Applicable'],6) #NA + Not Applicable

job_us.head()
```

Required Experience

```
#replacing and cleaning all the values in the "employment_type" into numeric values

job_us['employment_type'] = job_us['employment_type'].replace(['Full-time'], 0) #Full-time = 0
job_us['employment_type'] = job_us['employment_type'].replace(['Part-time'], 1) #Part-time = 1
job_us['employment_type'] = job_us['employment_type'].replace(['Contract'], 2) #Contract = 2
job_us['employment_type'] = job_us['employment_type'].replace(['Temporary'], 3) #Temporary = 3
job_us['employment_type'] = job_us['employment_type'].replace(['Other','NA'],4) #Other + NA = 4

job_us.head()
```

Employment Type

```
#replacing and cleaning all the values in "function" into numeric values divided in accordance to STEM

#Different types of "function" has been newly allocated into whether a candidate applied for a STEM related job or not
stem = ['Information Technology', 'Engineering', 'Data Analyst', 'Science', 'Business Analyst', 'Quality Assurance',
        'Manufacturing','Research']
non_stem = ['Sales', 'Customer Service', 'Marketing', 'Administrative', 'Design', 'Human Resources',
            'Health Care Provider', 'Other', 'Production', 'Education','Supply Chain', 'Business Development', 'Financi
            'Strategy/Planning', 'General Business', 'Product Management', 'Accounting/Auditing', 'Consulting', 'Financ
            'Advertising', 'Art/Creative','Public Relations','Training', 'Distribution', 'Purchasing','Legal','Project

job_us['function'] = job_us["function"].replace(stem, 1) #STEM = 1
job_us['function'] = job_us["function"].replace(non_stem, 0) #Non-STEM = 0
```

Function - STEM

```
#replacing and cleaning all the values in "required_education" into numeric values divided in accordance to DEGREE

#Different types of "required education" has been newly allocated into whether a candidate has a degree or not
Degree = ['Bachelor\'s Degree', 'Master\'s Degree', 'Associate Degree', 'Doctorate', 'Professional', 'Vocational - Degr
Non_Degree = ['NA', 'High School or equivalent', 'Unspecified', 'Some College Coursework Completed', 'Certification', '

job_us['required_education'] = job_us["required_education"].replace(Degree, 1) #Degree = 1
job_us['required_education'] = job_us["required_education"].replace(Non_Degree, 0) #Non_Degree = 0
job_us.head()
```

Required Education - Degree

Data Analysis (Text) - Statistical View

```
job_us["text"].iloc[0] #printing out the total infromation given in the first row(job)
```

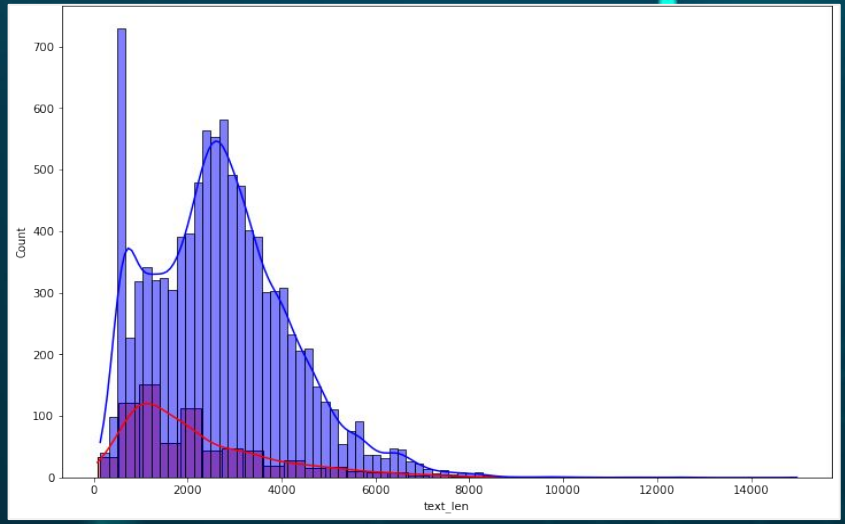
```
" Marketing Intern Marketing NA We're Food52, and we've created a groundbreaking and award-winning cooking site. We s
upport, connect, and celebrate home cooks, and give them everything they need in one place.We have a top editorial, b
usiness, and engineering team. We're focused on using technology to find new and better ways to connect people around
their specific food interests, and to offer them superb, highly curated information about food and cooking. We attrac
t the most talented home cooks and contributors in the country; we also publish well-known professionals like Mario B
atali, Gwyneth Paltrow, and Danny Meyer. And we have partnerships with Whole Foods Market and Random House.Food52 has
been named the best food website by the James Beard Foundation and IACP, and has been featured in the New York Times,
NPR, Pando Daily, TechCrunch, and on the Today Show.We're located in Chelsea, in New York City. Food52, a fast-growin
g, James Beard Award-winning online food community and crowd-sourced and curated recipe hub, is currently interviewin
g full- and part-time unpaid interns to work in a small team of editors, executives, and developers in its New York C
ity headquarters.Reproducing and/or repackaging existing Food52 content for a number of partner sites, such as Huffin
gton Post, Yahoo, Buzzfeed, and more in their various content management systemsResearching blogs and websites for th
e Provisions by Food52 Affiliate ProgramAssisting in day-to-day affiliate program support, such as screening affiliat
es and assisting in any affiliate inquiriesSupporting with PR &amp;amp; Events when neededHelping with office administrat
ive work, such as filing, mailing, and preparing for meetingsWorking with developers to document bugs and suggest imp
rovements to the siteSupporting the marketing and executive staff Experience with content management systems a major
plus (any blogging counts!)Familiar with the Food52 editorial voice and aestheticLoves food, appreciates the importan
ce of home cooking and cooking with the seasonsMeticulous editor, perfectionist, obsessive attention to detail, madde
ned by typos and broken links, delighted by finding and fixing themCheerful under pressureExcellent communication ski
llsA+ multi-tasker and juggler of responsibilities big and smallInterested in and engaged with social media like Twit
ter, Facebook, and PinterestLoves problem-solving and collaborating to drive Food52 forwardThinks big picture but pit
ches in on the nitty gritty of running a small company (dishes, shopping, administrative support)Comfortable with the
realities of working for a startup: being on call on evenings and weekends, and working long hours NA NA US"
```

Real Job Post Information

```
job_us["text"].iloc[101] #printing out the total infromation given in the first row(job)
```

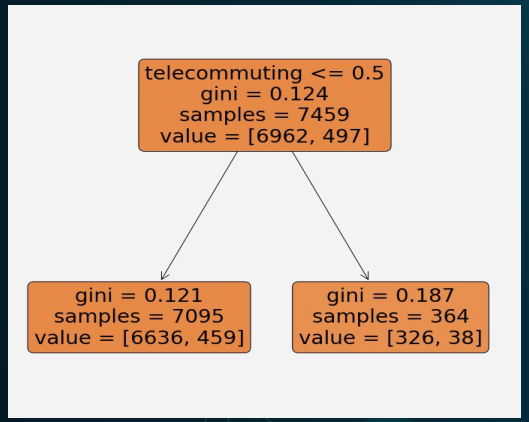
```
' CNC Programmer NA NA We Provide Full Time Permanent Positions for many medium to large US companies. We are interes
ted in finding/recruiting high quality candidates in IT, Engineering, Manufacturing and other highly technical and no
n-technical jobs. (We have more than 1500+ Job openings in our website and some of them are relevant to this job. Fee
l free to search it in the website and apply directly. Just Click the "Apply Now" and you will redirect to our main w
ebsite where you can search for the other jobs.)Job Requirements:Must be familiar with Job Shop type operations.CAM a
nd CAD experience a major plus.Ideal candidate will have a minimum of 10 yrs experience and have as strong of a manua
l manufacturing background as he does with CNC equipment. The machinery list for the facility is split between very l
arge CNC Mill, Manual Mills and Lathes and some small MAZAK (w/ Mazatrol Controls)Job Responsibilities:The Shift is 1
stThere is overtime, but it fluctuatesVisit - #URL_ec64af2b4fe2ca316e828f93b0cd098c22f8beba98dcac09d4dd7384b221a5e8#-
#URL_2954b76adf23051d4bc35fc729f5a18c5de715ed82dfaee24b3aabbf3874ca57# NA NA NA US'
```

Fake Job Post Information



Blue (Real): approx. 2500 words (in average)
Red (Fake): approx. 1600 words (in average)
Fake Jobs have less information provided!!!

Machine Learning - Decision Trees

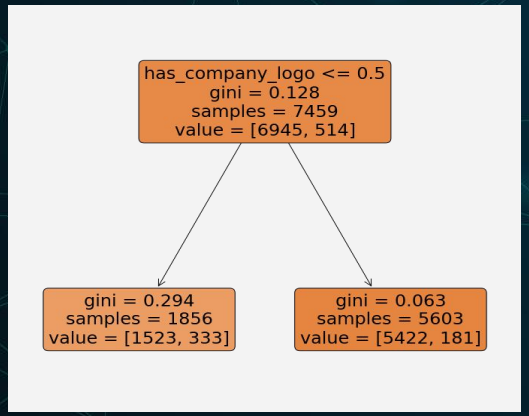



```
classesdecision(job_us, "telecommuting", "fraudulent",depth = 20)
```

```
predictor:        telecommuting
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9333690843276579

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9271191742258367
```

(Telecommuting = 0) 6.5% of Fake Job
(Telecommuting = 1) 10.4% of Fake Job
Fake Jobs use Telecommuting



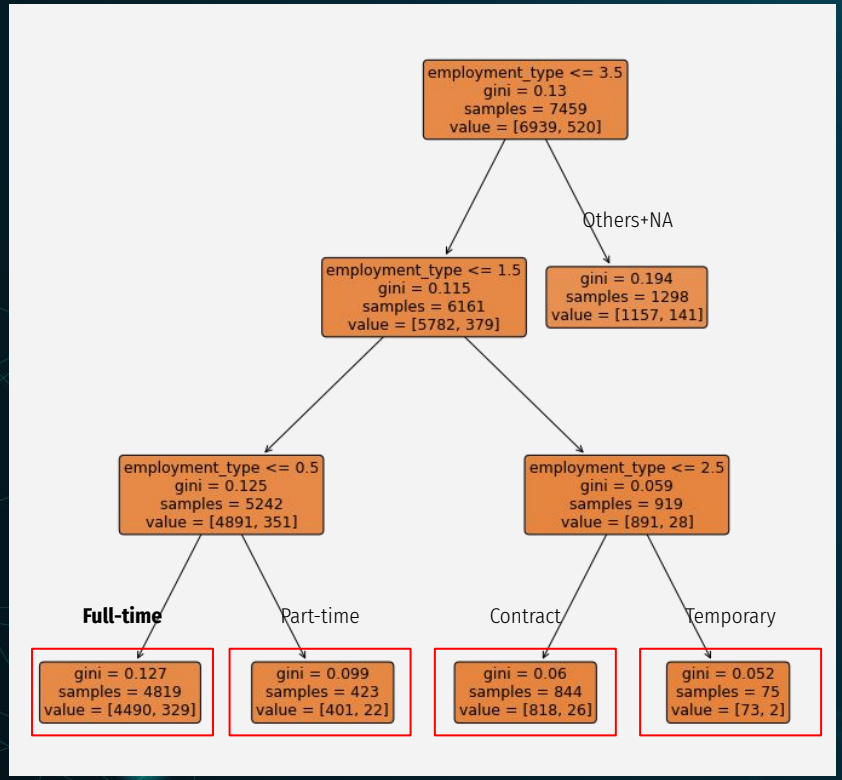

```
classesdecision(job_us, "has_company_logo", "fraudulent",depth = 20)
```

```
predictor:        has_company_logo
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9310899584394691

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9324366593681577
```

(Company Logo = 0) 17.9% of Fake Job
(Company Logo= 1) 3.2% of Fake Job
Fake Jobs does not have Logo

Machine Learning - Decision Trees



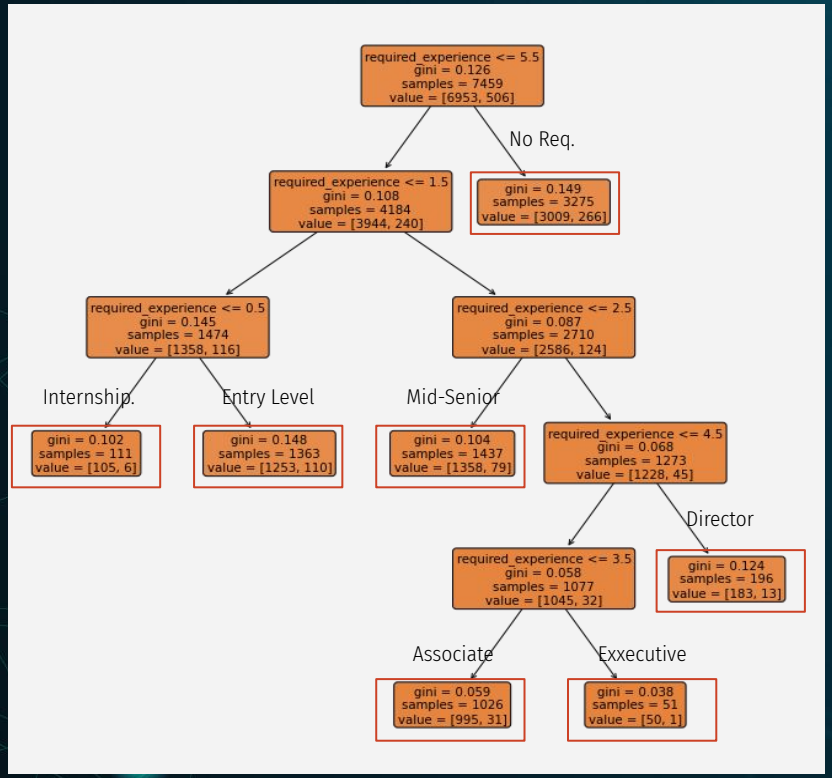

```
classesdecision(job_us, "employment_type", "fraudulent",depth = 20)
```

```
predictor:        employment_type
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9302855610671672

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9343134188301533
```

(Full-time = 0) 6.8% of Fake Jobs
(Part-time = 1) 5.2% of Fake Jobs
(Contract = 2) 3.1% of Fake Jobs
(Temporary = 3) 2.7% of Fake Jobs
Fake Jobs target Full-time candidates

Machine Learning - Decision Trees



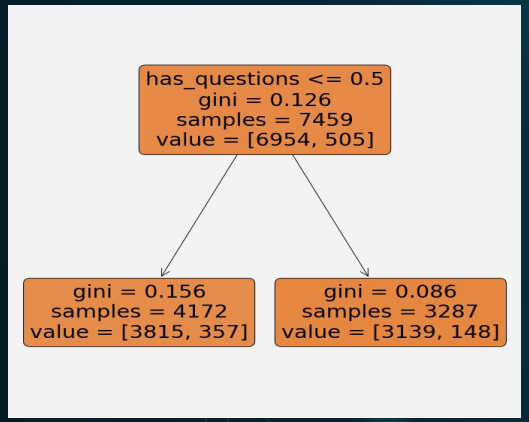

```
classesdecision(job_us, "required_experience", "fraudulent",depth = 20)
```

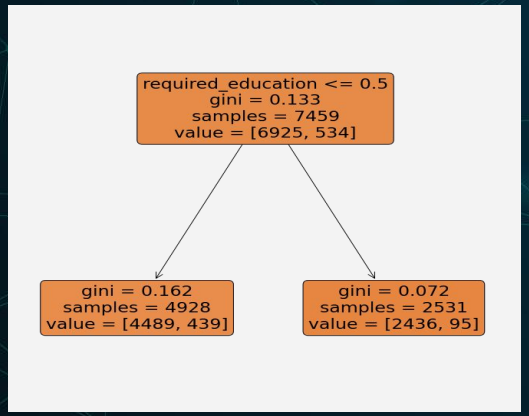
```
predictor:        required_experience
response:         fraudulent
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.932162488269205

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.9299343134188301
```

(Internship = 0) 5.4% of Fake Jobs
(Entry Level= 1) 8.1% of Fake Jobs
(Mid-Senior Level = 2) 5.5% of Fake Jobs
(Associate = 3) 3% of Fake Jobs
(Executive = 4) 2% of Fake Jobs
(Director = 5) 6.6% of Fake Jobs
(No Requirement = 6) 8.1% of Fake Jobs
Fake Job require entry level position or has no requirement

Machine Learning - Decision Trees






```
classesdecision(job_us, "has_questions", "fraudulent",depth = 20)
```

```
predictor:        has_questions
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9322965544979219

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9296215201751642
```

(questions = 0) 8.6% of Fake Job
(questions = 1) 4.5% of Fake Job
Fake Jobs does not ask questions

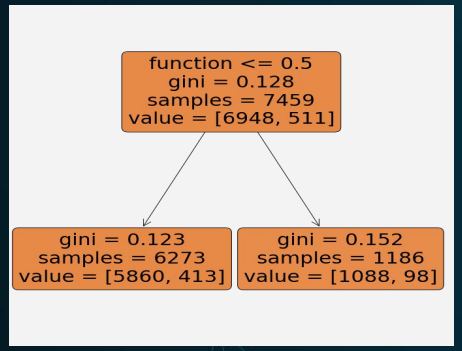
```
classesdecision(job_us, "required_education", "fraudulent",depth = 20)
```

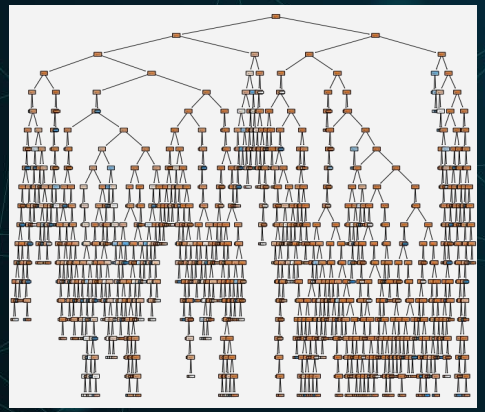
```
predictor:        required_education
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9284086338651294

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9386925242414764
```

(Degree= 1) 3.8% of Fake Job
(Non-Degree= 0) 8.9% of Fake Job
Fake Jobs does not require Degree

Machine Learning - Decision Trees





```
classesdecision(job_us, "function", "fraudulent",depth = 20)
```

```
predictor:        function
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9314921571256201

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.9314982796371598
```

(STEM = 0) 6.6% of Fake Job
(STEM = 1) 9.0% of Fake Job
Fake Jobs are usually STEM related

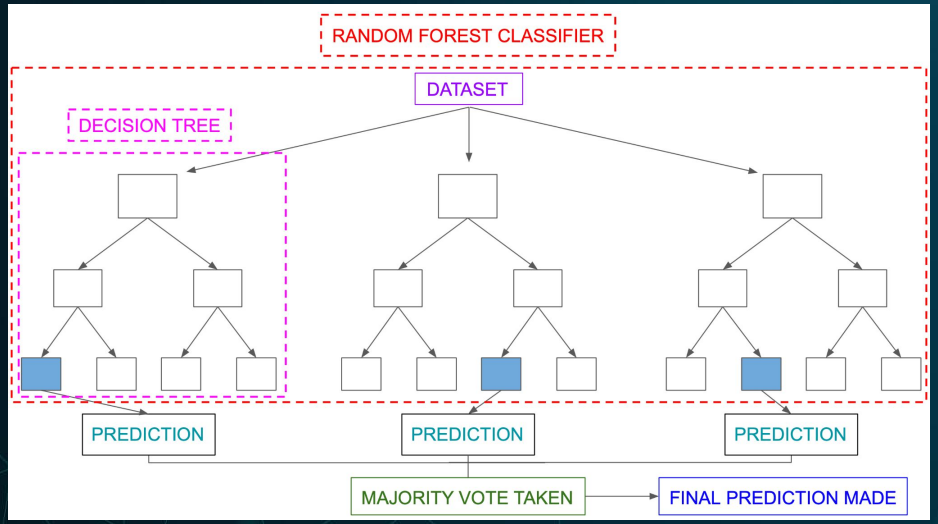
```
classesdecision(job_us, "text_len", "fraudulent",depth = 20)
```

```
predictor:        text_len
response:         fraudulent
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.9621933235018099

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.920863309352518
```

Text length is prone to overfitting due to
High variance and is less biased.
It needs Random Forest Classification

Machine Learning - Random Forest




```
def ForClass(data, predictor, response, test_ratio, model = RandomForestClassifier()):
    print("Predictor:\t", predictor)
    print("Response:\t", response)
    X = pd.DataFrame(data[predictor])
    y = pd.DataFrame(data[response])

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = test_ratio)

    dectree = RandomForestClassifier(n_estimators=20, random_state = 42)
    dectree.fit(X_train, y_train.values.ravel())
    y_pred = dectree.predict(X_test)
```

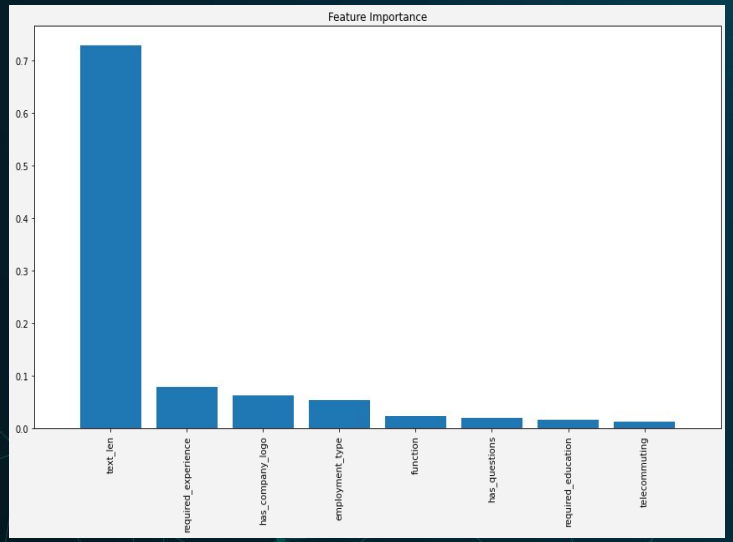
```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.9939670197077356

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.9321238661244917
```

1. Prevents Overfitting
2. Higher Accuracy and Precision

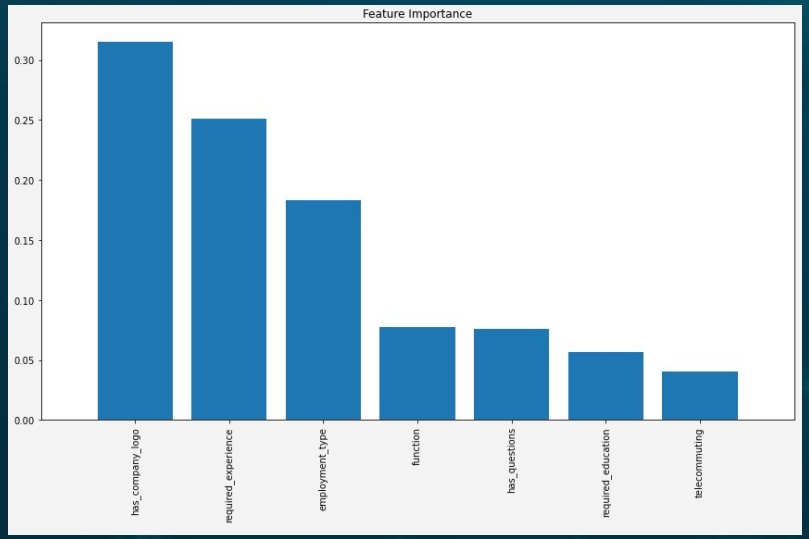
Machine Learning - Random Forest

Value Importance



1)	text_len	0.729553
2)	required_experience	0.079101
3)	has_company_logo	0.063300
4)	employment_type	0.054088
5)	function	0.024506
6)	has_questions	0.020763
7)	required_education	0.016000
8)	telecommuting	0.012689

Value Importance (excluding text_len)



1)	has_company_logo	0.315178
2)	required_experience	0.251327
3)	employment_type	0.182820
4)	function	0.077354
5)	has_questions	0.076130
6)	required_education	0.056506
7)	telecommuting	0.040686

NLP Approach

03

Fake job detection by comparing different machine learning models

Data transformation

Loading the English tokenizer, tagger, parser and word vectors

Changing words into vectors using TFID vectorizer

Creating list of punctuation marks, stop words

Customized transformer using spaCy

Fit to Model, Test_size = 0.25

Logistic Regression

Fit an "S" shaped logistic function, which predicts two maximum values (Real Job or Fake Job).

from sklearn.linear_model import LogisticRegression

```
Logistic Regression Accuracy: 0.9652096342551294
Logistic Regression Recall: 0.0
```

Support vector machine

Find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

from sklearn.svm import SVC

```
SVC Accuracy: 0.9652096342551294
SVC Recall: 0.0
```

Fit to Model, Test_size = 0.25

Random Forest

Ensemble classifier that estimates based on the combination of different decision trees

from sklearn.ensemble import RandomForestClassifier

```
Random Forest Accuracy: 0.9937555753791257
Random Forest Recall: 0.8205128205128205
```

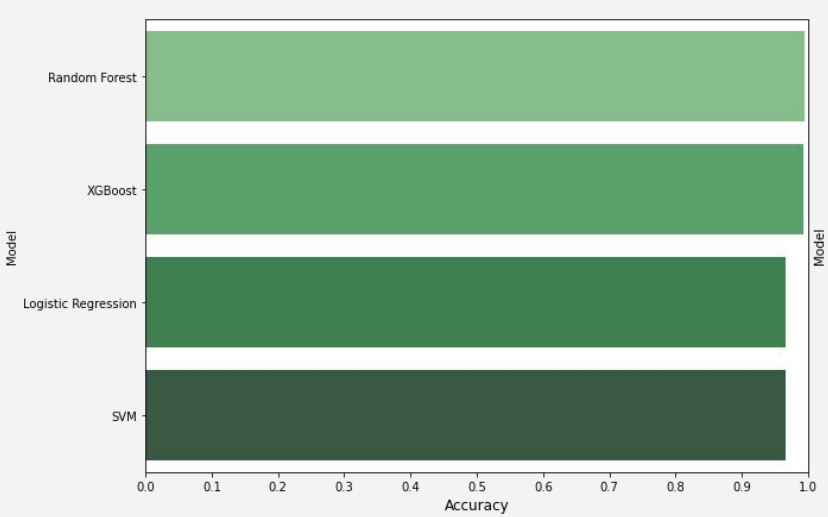
XGBoost

XGBoost is decision-tree-based ensemble Machine Learning algorithm that uses gradient boosting framework

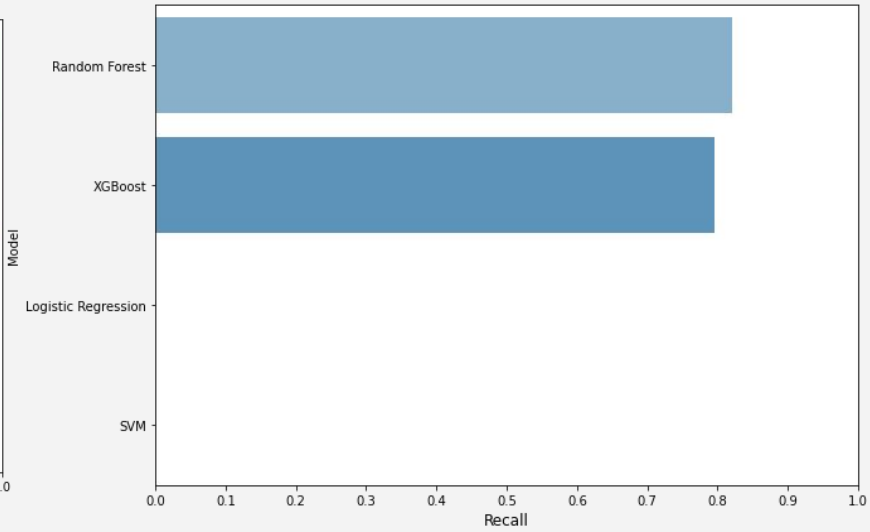
from xgboost import XGBClassifier

```
XGBoost Accuracy: 0.9928635147190009
XGBoost Recall: 0.7948717948717948
```

Comparison of models

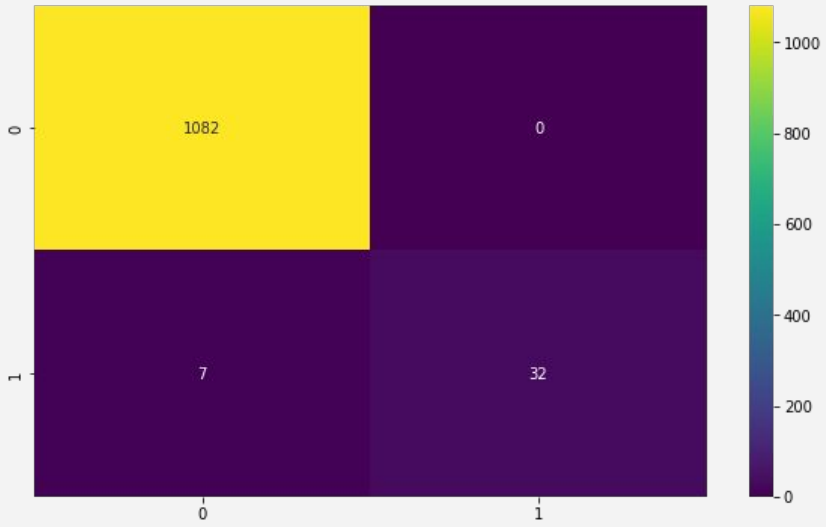


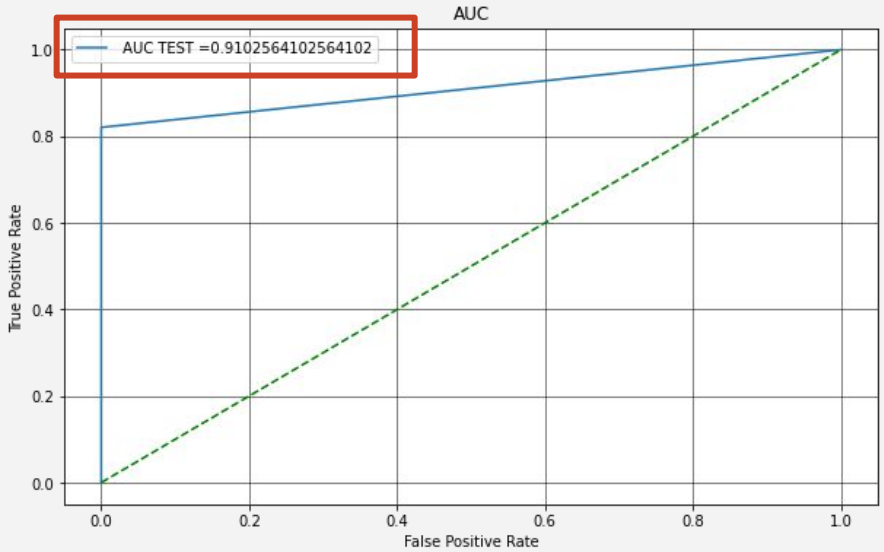
Accuracy



Recall

Best Model Evaluation

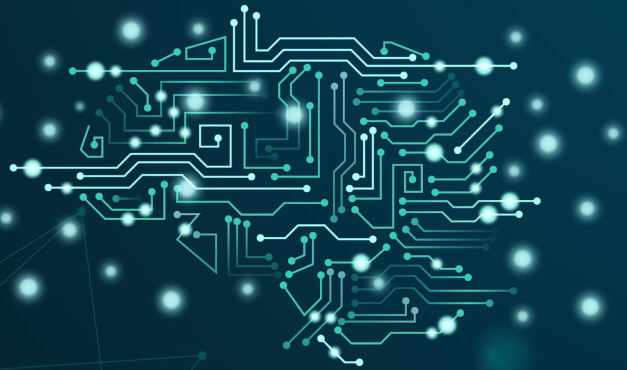




Using random forest approach for this dataset show a high level of accuracy

Thank you

Team Koviema



Quote from “Koviema”:

Change is the end result of all true learning